The Importance of Labeling and Noise on The Trading Strategy

Zinnet Duygu Akşehir
Department of Computer Engineering
Ondokuz Mayıs University
Samsun, Türkiye
duygu.aksehir@bil.omu.edu.tr

Erdal Kılıç
Department of Computer Engineering
Ondokuz Mayıs University
Samsun, Türkiye
erdal.kilic@bil.omu.edu.tr

2023 8th International Conference on Computer Science and Engineering (UBMK) | 979-8-3503-4081-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/UBMK59864.2023.10286793

***Abstract*—This study proposed an indirectly developed trading strategy based on the labeling methods used in the literature on noiseless financial time series using the CEEMDAN method. This proposed method shows the importance of labeling and noise to increase the profits the stock market investor will obtain from the stock trade. The efficiency of the developed method was tested on AAPL, AMZN, INTC, NVDA, THYAO.IS and GARAN.IS stocks. The results obtained prove the effectiveness of the method.**

Keywords— stock; CEEMDAN; labeling, financial time series; noise.

I. Introduction

Stock market forecasting, located at the intersection of finance and computer technology, is among the most interesting topics and attracts great attention from researchers and investors. The main reason for this interest is the investors' expectation of making high profits.

Studies within the scope of stock market forecasting have generally focused on predicting the direction of a commodity (stock, stock market index, gold, etc.). These studies are carried out using supervised learning techniques. Therefore, before the training process, the data must be labeled. For this, a labeling approach is needed. When we examine studies in the literature, the up-down labeling approach [1]–[3] is generally used, but there are also window-based labeling approaches [4], [5]. Although window-based labeling approaches were proposed to overcome the disadvantage of the up-down approach being affected by small price changes, they also have several disadvantages. For example, in these approaches, there is no information about choosing the window size, but the labels also vary according to this window size. This situation directly affects the model performance, leading to decreased profitability. In addition, another problem in the developed models is that the prediction models are sensitive to noise. This is because the stock market is influenced by many factors, such as company policies, political events, investor expectations, and the general economic situation [6]. These factors cause the stock market to move independently at every point, and financial time series contain significant noise [7]. Using noisy financial time series directly in forecasting models will lead to the development of noise-sensitive models. For this reason, when developing models, there will be a problem of mislabeling due to data containing noise, reducing the model's prediction success. This situation will cause the investor to make wrong decisions and, therefore, be unable to get the expected profit.

When the studies that eliminate noise in financial time series are examined, it is seen that noise removal approaches based on Fourier transform, wavelet transform, and signal decomposition methods are preferred to improve the model's prediction performance. Song et al. [8] proposed a padding-based Fourier transform approach to predict the closing prices of the S&P500, KOSPI, and SSE indices to eliminate noisy components in the index data that negatively affect the model performance. Combining the proposed denoising technique with deep learning methods was found to outperform the basic models. As a result of the experiments, it was stated that the hybrid model consisting of the denoising technique and deep learning methods was more successful than the models that did not apply the denoising technique. Because the Fourier transform was less effective in analyzing time-varying signals, wavelet transform-based denoising techniques were used. This approach eliminated noisy components in daily and minute stock market data [9]–[11]. Dastgerdi and Mercorell [9] stated that it is a challenging problem to create a forecasting model on these data due to the characteristics of financial time series and pointed out that it is inevitable to use denoising techniques to eliminate the noise in stock data. Accordingly, five different stock market indices were selected to evaluate the effect of the Kalman filter and wavelet transform denoising techniques on the prediction model. As a result of the experiments, it was seen that the denoising techniques used are an important factor in increasing the model's prediction performance. Although the wavelet transform method effectively eliminates noise in financial time series, according to the studies examined, the effectiveness of this method depends on some parameters. These parameters are the number of decomposition layers and the essential wavelet function. There needs to be definite information about these parameters to eliminate the noise in financial time series effectively. Therefore, these parameters are usually decided by trial and error in studies. To overcome all these situations, mode decomposition-based denoising methods came to the fore recently. In this context, empirical mode decomposition

(EMD), ensemble empirical mode decomposition (EEMD), complete ensemble empirical mode decomposition (CEEMD), and complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) were frequently preferred [12]–[17]. When the studies in the literature were examined, it was proven that these mode decomposition techniques are more successful than the methods that do not use denoising techniques [12]–[16]. When these hybrid methods based on mode decomposition used to predict stock market closing prices in the literature were examined, they were applied to indices such as S&P500, HSI, DAX, SSE, and DJIA. In all these studies, it was seen that the CEEMDAN method is preferred more than other mode decomposition methods because the CEEMDAN method provides better mode decomposition thanks to adaptive noise control.

The primary motivation of this study is to develop an indirect solution proposal that will provide this situation instead of creating a new labeling approach that will increase the profitability of investors. Considering the labeling approaches used, the main problems in these studies are the sensitivity to small price data and the uncertainty in the choice of window size. In this direction, the main contribution of this study is to implement labeling approaches using the data obtained as a result of eliminating the noise in these data by a CEEMDAN-based method, as opposed to performing them directly on closing prices that contain noise. Thus, it will be possible for the investor to make more profit compared to the first situation.

The remaining parts of the study are organized as follows: Section II mentions labeling approaches and the CEEMDAN-based denoising method. The experimental results from the study are detailed in Section III. The last section consists of the conclusion and future works.

II. Method

The labeling approaches used in the literature and the CEEMDAN-based denoising method are mentioned in this section.

A. Labeling Approaches

When the stock market forecasting studies are examined in recent years in literature, although the results seem satisfactory, it is determined that the labeling algorithms used in these studies trigger some problems. In this context, three different labeling algorithms used in the literature and the problems they cause are listed below:

1. *Up-Down Labeling Approach:* While labeling according to this method, the closing value of two consecutive trading days is considered, and labeling is carried out according to Equation 1. Here t is the trading day, C is the closing value, and L is the label.

$$L_{t+1} = \begin{cases} Up, & C_t < C_{t+1} \\ Down, & C_t \geq C_{t+1} \end{cases} \quad (1)$$

2. *Sezer and Özbayoğlu's Labeling Approach [5]:* It is one of the proposed window-based labeling approaches to overcome the disadvantage of the up-down approach, which is sensitive to small price changes. In this proposed algorithm, the 11-day sliding window approach is used. In this approach, labeling is performed according to the midpoint of the 11-day window, in other words, the closing price of the 6th day. If the closing price of the 6th day is the maximum value of this 11-day window, it is labeled "Down"; if the minimum value is "Up". This proposed labeling approach is an effective method for labeling V and inverted-V-shaped patterns.
3. *NPMM Labeling Approach [4]:* It is another of the proposed window-based labeling approaches to overcome the disadvantage of the up-down approach. This situation, which the up-down labeling approach has, negatively affects model training. In this direction, a new labeling algorithm called NPMM (N-period min-max labeling) is proposed, which defines the trend for N periods and labels their minimum-maximum values. This labeling approach is carried out with Equation 2. The N_{max} and N_{min} specified here represent the maximum and minimum closing values for the N period.

$$L_{t+1} = \begin{cases} Up, & C_t == N_{min} \\ Down, & C_t == N_{max} \end{cases} \quad (2)$$

The main problems with these labeling approaches are:

- Since two consecutive days are considered in the up-down approach, this approach is sensitive to minor price fluctuations. Therefore, labeling stock prices with high noise in this way will cause false up-down signals, causing the investor to make a loss.
- In window-based labeling approaches, it is still unclear what the ideal window size is to choose. As the window size changes, so does the success of the labeling. For example, in the NPMM labeling approach, if the N window size is selected as small, the disadvantage in the up-down approach is encountered. In contrast, if it is set large, less labeling is performed, and this causes the investor to make less profit than expected.

B. CEEMDAN-based Denoising Technique

The EMD method proposed by Huang et al. divides a time series into essential components with different frequencies and scales, called intrinsic mode function (IMF) [18]. It can decompose complex nonlinear and non-stationary time series data into multiple IMF components. The last of these IMFs is also called the residue. These obtained IMFs can be used in data analysis applications or to understand the internal structure of a time series. But this method, in practice, leads to the problem of mode mixing, which refers to the presence of very similar oscillations in different modes or very different amplitudes in one mode. To overcome this disadvantage of the EMD method, EEMD is proposed as an improved version [19]. The EEMD defines the basic IMF components by averaging over several trials. It is also a noise-assisted decomposition method that adds white noise to each trial and

thus enables obtaining less noisy IMFs than EMD. Therefore, since EEMD performs noise-assisted decomposition and is an ensemble approach, it provides superiority to the EMD method and essentially eliminates the mode mixing problem. However, EEMD has a high computational cost and cannot wholly eliminate white noise after signal reconstruction. To solve these problems, Torres et al. proposed the CEEMDAN method as an enhanced version of the EEMD [20]. CEEMDAN is superior to other mode separation methods because it effectively eliminates the problem of mode mixing, the reconstruction error is almost zero, and the computational cost is significantly reduced. When a non-stationary and non-linear $X(t)$ signal is given, decomposition is performed with the CEEMDAN method by following the steps in Algorithm 1. In

Algorithm 1 CEEMDAN Decomposition Algorithm

Output: $\widetilde{IMF_k}$
$r_k(n) \leftarrow residues$
$E_j(.) \leftarrow j$-th IMF obtained by EMD decomposition
$w^i \leftarrow$ White noise
$x(n) \leftarrow$ Time series signal
$\epsilon_0 \leftarrow$ Noise coefficient
$I \leftarrow$ Number of trials
$IMF_1^i(n) = x(n) + \epsilon_0 w^i(n)$
$\widetilde{IMF_1}(n) = 0$
for $i = 1$ to I **do**
  $\widetilde{IMF_1}(n) = \widetilde{IMF_1}(n) + IMF_1^i(n)/I \leftarrow$ First IM
end for
$r_1(n) = x(n) - \widetilde{IMF_1}(n) \leftarrow$ First residue
$\widetilde{IMF_2}(n) = 0$
for $i = 1$ to I **do**
  $\widetilde{IMF_2}(n) = \widetilde{IMF_2}(n) + E_1(r_1(n) + \epsilon_1 E_1(w^i(n)))/I \leftarrow$ Second IMF
end for
while $r_k(n) \leftarrow$ until the value of residual component is less than two extremes **do**
  for $k = 2$ to K **do**
    $r_k(n) = r_{k-1}(n) - \widetilde{IMF_k}(n) \leftarrow$ Residuals for $k = 1, 2, ..K$
    $\widetilde{IMF_k}(n) = 0$
    for $i = 1$ to I **do**
      $\widetilde{IMF_{k+1}}(n) = \widetilde{IMF_{k+1}}(n) + E_1(r_k(n) + \epsilon_k E_k(w^i(n)))/I$
    end for
  end for
end while

separating noisy components from IMFs obtained as a result of decomposition, the first two components with high frequency are usually discarded [16]. In this study, these high-frequency components were discarded, and the remaining components were summed to obtain noiseless time series data. Based on the labeling mentioned above algorithms on this noiseless data obtained, the data were first labeled as up-down. Then, buy-sell signals are created by applying the following steps to these labeled data:

- When the Up label is encountered, a "Buy" signal is generated, and the Up label's equivalent is determined as "Hold" until the Down label is encountered.
- After the" Buy "signal is created, the" Sell "signal is generated for the equivalent of the first encountered Down label and the equivalent of the days until the next Up label is determined as" None."

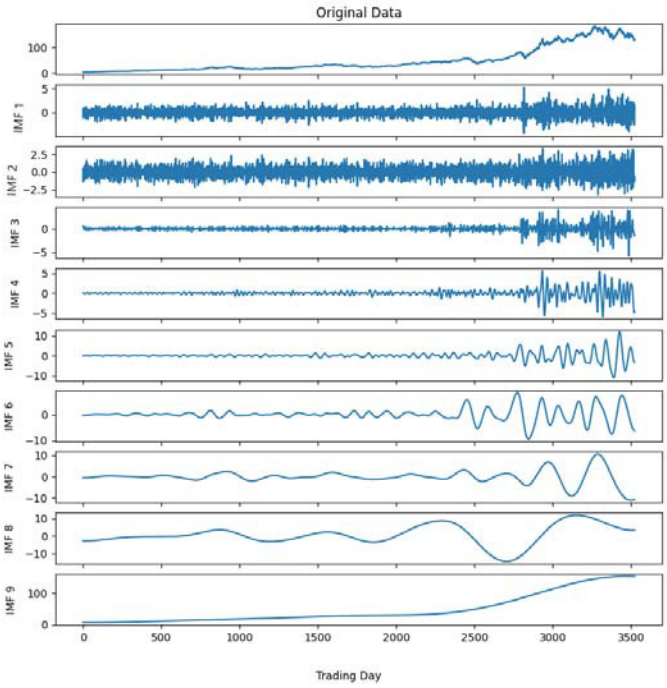

Fig. 1. Decomposition results of AAPL stock.

III. Experimental Results

Various experiments were carried out to evaluate the effectiveness of the CEEMDAN-based denoising method for labeling financial time series. In this direction, to enable the investor to make more profit, the labeling process on the noiseless data was applied to a large number of stocks, and due to the page limit in the paper, only the findings of AAPL stock were detailed.

Firstly, the daily closing values for APPL stock between January 1, 2009, and December 31, 2022, were obtained from Yahoo! Finance. Then, the CEEMDAN method was applied to these data, and the obtained IMFs were shown in Figure 1. Accordingly, due to the high frequency of the first two IMFs, these IMFs were discarded, and created noiseless data from the remaining IMFs. Then, using the three labeling approaches described in the Section II-A on the noiseless data, first labeling as Up-Down was performed, and then generated Buy-Sell signals over these labels. These signals, obtained from the original and denoised closing price data of the AAPL stock, are given in Figure 2-4. In addition, the profit obtained with the

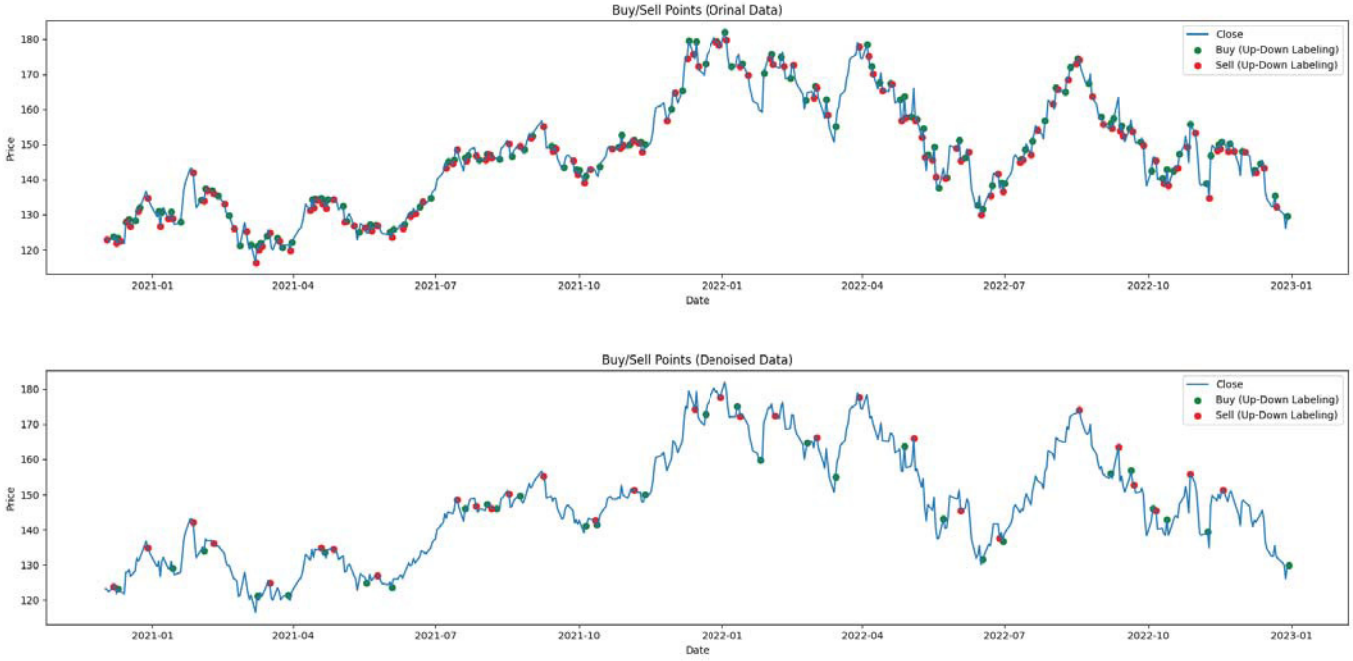

Fig. 2. The last two years of trading signals with obtained the up-down labeling approach.

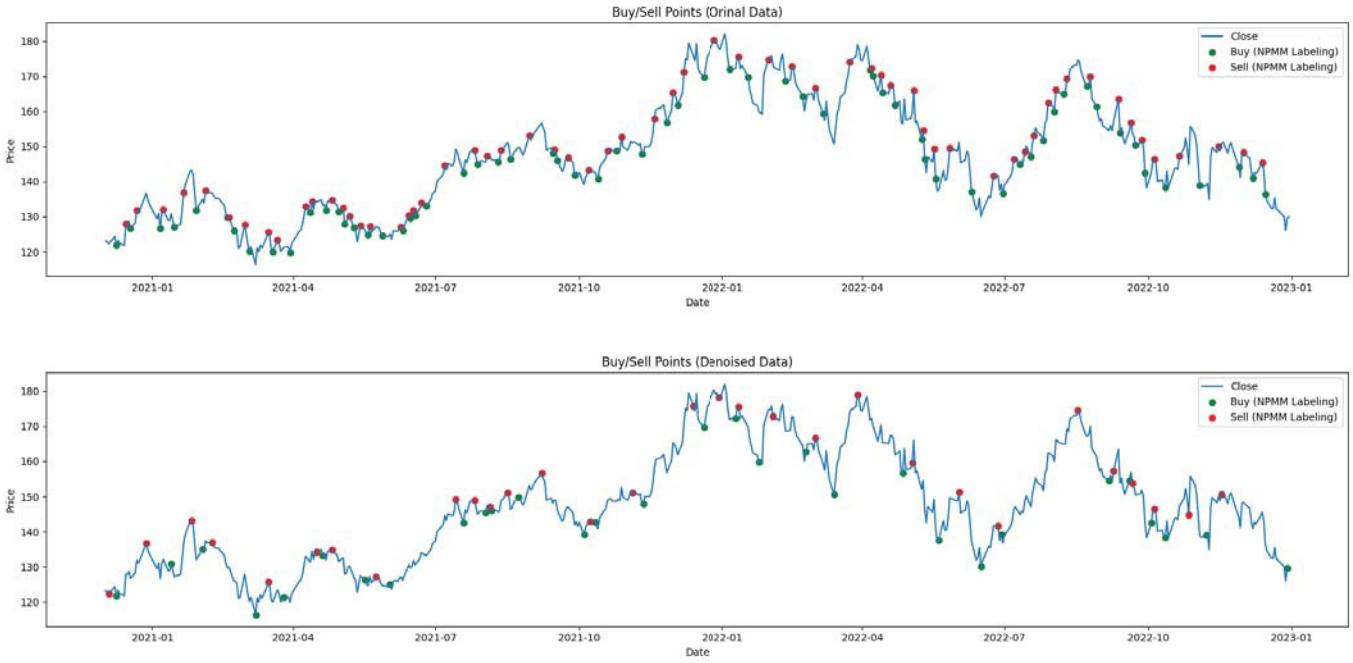

Fig. 3. The last two years of trading signals with obtained the NPMM labeling approach.

proposed method was calculated according to the procedure in Algorithm 2. It was determined with 1000 $ as initial capital, and a commission fee of 0.002 was deducted from each trading transaction. While calculating the profitability, carried out Buy-Sell transactions over the original opening and closing values. The profit obtained from the trading operations performed for six stocks is given in Table I.

In addition, applied similar procedures were used in AMZN,

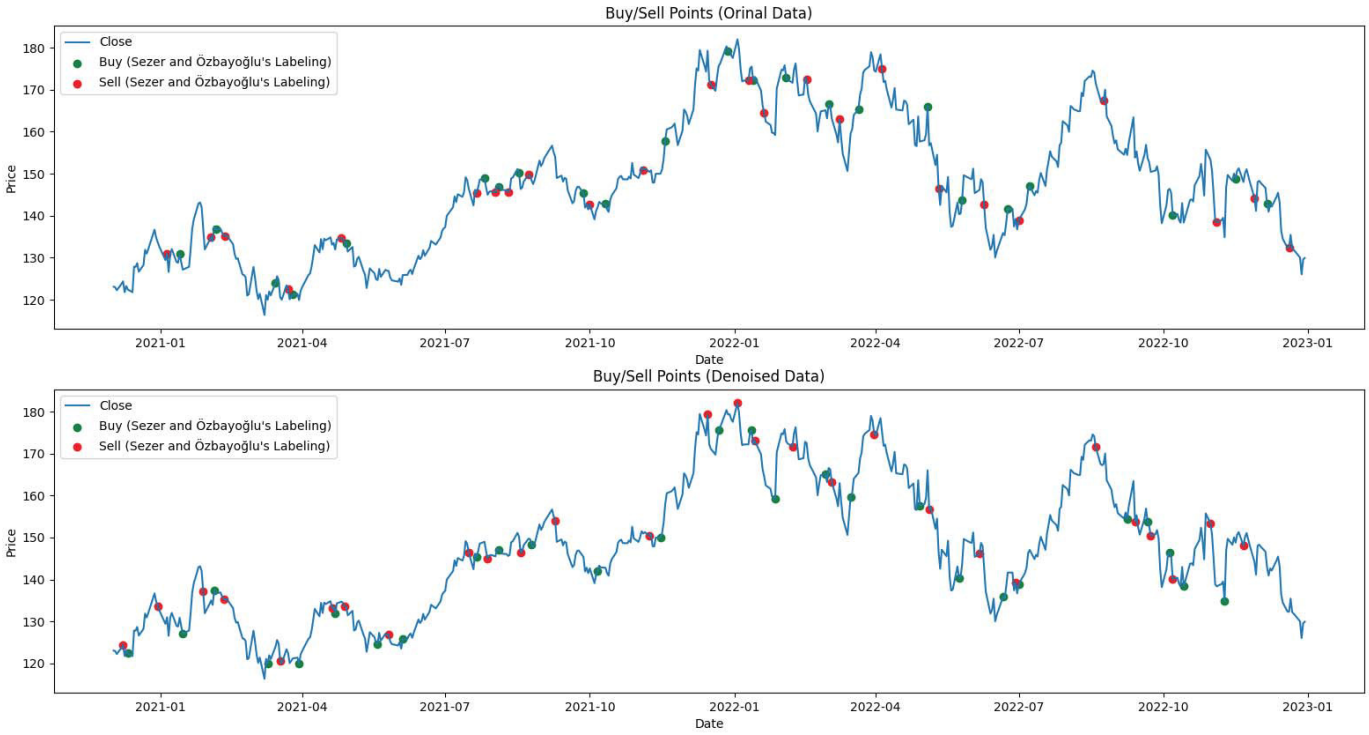

Fig. 4. The last two years of trading signals with obtained Sezer and Özbayoğlu's labeling approach.

TABLE I
THE PROFIT OBTAINED AS A RESULT OF TRADING ACCORDING TO THE LABELING APPROACHES OF STOCKS.

Stocks		Up-Down Labeling	NPMM Labeling	Sezer and Özbayoğlu's Labeling
AAPL	Original Data	7180.888 $	8110.893 $	2990.418 $
	Denoised Data	**10975.579 $**	**11509.901 $**	**8729.017 $**
AMZN	Original Data	8274.394 $	9525.533 $	1789.707 $
	Denoised Data	**11487.679 $**	**12331.96 $**	**9211.412 $**
INTC	Original Data	4355.06 $	6513.664 $	-207.815 $
	Denoised Data	**10656.132 $**	**10665.366 $**	**8531.546 $**
NVDA	Original Data	12250.018 $	12599.148 $	3320.279 $
	Denoised Data	**13693.241 $**	**13350.222 $**	**10179.751 $**
THYAO.IS	Original Data	9379.193 $	13591.093 $	5891.645 $
	Denoised Data	**16347.894 $**	**16963.221 $**	**12606.655 $**
GARAN.IS	Original Data	6266.897 $	11580.547 $	872.461 $
	Denoised Data	**16059.125 $**	**16708.332 $**	**11929.755 $**

INTC, NVDA, THYAO.IS and GARAN.IS stocks, and the profit results obtained were also shared in Table I. When the results were examined, it was seen that more profit was obtained for all three labeling approaches as a result of realization of the buy-sell signals produced on noiseless data. In the labeling approaches of NPMM and Sezer-Özbayoğlu, window size 11 was selected for the original data, while window sizes 2 and 5 were chosen for noiseless data, respectively. In these approaches, the window size was set as 11 so that noise during labeling would not affect the data. Since the data used in our proposed method is already noiseless, a smaller window size was sufficient instead of choosing a size of this size.

IV. CONCLUSION AND FUTURE WORKS

This study presents an indirect solution proposal instead of developing a new labeling approach that will increase the investor's profitability. Considering the labeling approaches used, there are several problems, such as sensitivity to small price data and uncertainty in the choice of window size. In this direction, labeling approaches were carried out on the noiseless data obtained from applying the CEEMDAN-based denoising method on the closing prices instead of performing directly on the closing prices. With this proposed approach, it is shown that investors can make more profit by taking into the noise and labeling correctly.

Developing a more effective denoising technique for financial time series is considered a future work.

Algorithm 2 Calculation of Profitability

procedure CALCULATE_PROFIT($Data, Label$)
  $comission\_rate = 0.002$
  $buy\_points = Data[Data[Label] == "Buy"].index$
  $sell\_points = Data[Data[Label] == "Sell"].index$
  $total\_profit = 0$
  for i **in range** (len(buy_points)) **do**
    $starting\_capital = 1000\$$
    $capital = starting\_capital$
    $buy\_price = Data["Open"][i]$
    $lots = capital // buy_price$
    $capital- = ((lots \times buy\_price) \times (1 + comission\_rate))$
    $sell\_index = sell\_points[sell\_points > buy\_index].min()$
    if notnull(sell_index) **then**
      $sell\_price = Data["Close"][sell\_index]$
      $capital+ = ((lots \times sell\_price) \times (1 - comission\_rate))$
      $profit = capital - starting\_capital$
      $total\_profit+ = profit$
    end if
  end for
  return $total\_profit$
end procedure

ACKNOWLEDGMENT

This work was supported by Ondokuz Mayıs University BAP under grant PYO.MUH.1904.23.002.

REFERENCES

[1] J. Shen and M. O. Shafiq, "Short-term stock market price trend prediction using a comprehensive deep learning system," *Journal of big Data*, vol. 7, no. 1, pp. 1–33, 2020.

[2] S. Basak, S. Kar, S. Saha, L. Khaidem, and S. R. Dey, "Predicting the direction of stock market prices using tree-based classifiers," *The North American Journal of Economics and Finance*, vol. 47, pp. 552–567, 2019.

[3] K. K. Yun, S. W. Yoon, and D. Won, "Prediction of stock price direction using a hybrid ga-xgboost algorithm with a three-stage feature engineering process," *Expert Systems with Applications*, vol. 186, p. 115716, 2021.

[4] Y. Han, J. Kim, and D. Enke, "A machine learning trading system for the stock market based on n-period min-max labeling using xgboost," *Expert Systems with Applications*, vol. 211, p. 118581, 2023.

[5] O. B. Sezer and A. M. Ozbayoglu, "Algorithmic financial trading with deep convolutional neural networks: Time series to image conversion approach," *Applied Soft Computing*, vol. 70, pp. 525–538, 2018.

[6] N. Rouf, M. B. Malik, T. Arif, S. Sharma, S. Singh, S. Aich, and H.-C. Kim, "Stock market prediction using machine learning techniques: a decade survey on methodologies, recent developments, and future directions," *Electronics*, vol. 10, no. 21, p. 2717, 2021.

[7] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *The journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[8] D. Song, A. M. C. Baek, and N. Kim, "Forecasting stock market indices using padding-based fourier transform denoising and time series deep learning models," *IEEE Access*, vol. 9, pp. 83 786–83 796, 2021.

[9] A. K. Dastgerdi and P. Mercorelli, "Investigating the effect of noise elimination on lstm models for financial markets prediction using kalman filter and wavelet transform," *WSEAS Trans. Bus. Econ*, vol. 19, pp. 432–441, 2022.

[10] Q. Tang, R. Shi, T. Fan, Y. Ma, and J. Huang, "Prediction of financial time series based on lstm using wavelet transform and singular spectrum analysis," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–13, 2021.

[11] W. Bao, J. Yue, and Y. Rao, "A deep learning framework for financial time series using stacked autoencoders and long-short term memory," *PloS one*, vol. 12, no. 7, p. e0180944, 2017.

[12] H. Rezaei, H. Faaljou, and G. Mansourfar, "Stock price prediction using deep learning and frequency decomposition," *Expert Systems with Applications*, vol. 169, p. 114332, 2021.

[13] J. Cao, Z. Li, and J. Li, "Financial time series forecasting model based on ceemdan and lstm," *Physica A: Statistical mechanics and its applications*, vol. 519, pp. 127–139, 2019.

[14] P. Lv, Q. Wu, J. Xu, and Y. Shu, "Stock index prediction based on time series decomposition and hybrid model," *Entropy*, vol. 24, no. 2, p. 146, 2022.

[15] B. Yan, M. Aasma *et al.*, "A novel deep learning framework: Prediction and analysis of financial time series using ceemd and lstm," *Expert systems with applications*, vol. 159, p. 113609, 2020.

[16] Y. Liu, X. Liu, Y. Zhang, and S. Li, "Cegh: A hybrid model using ceemd, entropy, gru, and history attention for intraday stock market forecasting," *Entropy*, vol. 25, no. 1, p. 71, 2023.

[17] C. Zhang, Q. Lan, X. Mi, Z. Zhou, C. Ma, and X. Mi, "A denoising method based on the nonlinear relationship between the target variable and input features," *Expert Systems with Applications*, vol. 218, p. 119585, 2023.

[18] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[19] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in adaptive data analysis*, vol. 1, no. 01, pp. 1–41, 2009.

[20] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, "A complete ensemble empirical mode decomposition with adaptive noise," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 4144–4147.