SHM PUBLISHER
SURYA HIJAU MANFAAT

Journal of Soft Computing Exploration

Homepage: https://shmpublisher.com/index.php/joscex

JOSCEX

p-ISSN: 2746-7686
e-ISSN: 2746-0991

.

Ensemble learning technique to improve breast cancer classification model

Ahmad Ubai Dullah[1], Fitri Noor Apsari[2], Jumanto[3]
[1,2,3]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received May 31, 2023
Revised Jun 26, 2023
Accepted Jun 29, 2023

Keywords:

Breast cancer
Classification
Decision tree
Logistic regression
Random forest
Ensemble learning

ABSTRACT

Cancer is a disease characterized by abnormal cell growth and is not contagious, such as breast cancer which can affect both men and women. breast cancer is one of the cancer diseases that is classified as dangerous and takes many victims. However, the biggest problem in this study is that the classification method is low and the resulting accuracy is less than optimal. the purpose of this study is to improve the accuracy of breast cancer classification. Therefore, a new method is proposed, namely ensemble learning which combines logistic regression, decision tree, and random forest methods, with a voting system. This system is useful for finding the best results on each parameter that will produce the best prediction accuracy. The prediction results from this method reached an accuracy of 98.24%. The resulting accuracy rate is more optimal by using the proposed method.

Corresponding Author:

Ahmad Ubai Dullah,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran, Gunungpati, Semarang, Indonesia.
Email: ubaid@students.unnes.ac.id
https://doi.org/10.52465/joscex.v4i2.166

1. INTRODUCTION

Cancer is a type of non-communicable disease characterized by the constant and malignant growth of abnormal cells or tissues that can affect the function of the affected tissues Cancer cells arise from various organ-forming elements and can form tumor masses through excessive cell division and spread through blood vessels or lymph nodes [1], [2]. One of the deadliest types of cancer is breast cancer. Breast cancer affects both men and women, but is more common in women and very rare in men [3], [4]. The level of heterogeneity in breast cancer is high. Breast cancer is currently the second cause of cancer deaths in women. Breast cancer can be divided into benign and malignant types [5]. Benign breast cancer is a non-invasive form of breast cancer that rarely endangers the patient's life. Benign breast cancers are found in the lining of the breast ducts and do not spread to the surrounding tissues [6].

According to an article in A Cancer Journal for Clinicians published by CA in 2020, there were 2.3 million women with breast cancer or about 11.7% of all newly diagnosed cancer cases. Breast cancer has a higher prevalence than lung cancer, with a prevalence of only 11.5%. Hyuna Song, a senior scientist and epidemiologist at the American Cancer Society said breast cancer is the most common of all cancers, with cases up 2.3 million from 2,088,849 in 2018 [7]. According to data from the Global Cancer Observatory (GLOBOCAN) in 2018, breast cancer is classified as dangerous cancer, ranking second out of five cancers with the highest number of patient deaths in Indonesia. Of the 207,210 total deaths, 11% or 22,692 died of breast cancer [8].

If you look at the results of previous research, the accuracy of the classification results is still not getting optimal accuracy. This study aims to improve the accuracy resulting from the classification of breast cancer and also be more optimal in predicting the level of breast cancer. Therefore, we propose a model, namely the Ensemble Learning [9] technique. This technique makes it possible to combine several research methods, the methods we use in this combination include decision tree, random forest, and logistic regression methods [10], [11]. we combined the three methods to produce more optimized results.

2. METHOD

The design flow of the proposed algorithm is depicted as shown in Figure 1. The algorithm used is Ensamble Learning which combines 3 research methods. Each process in Figure 1 will be explained in detail in the next section.

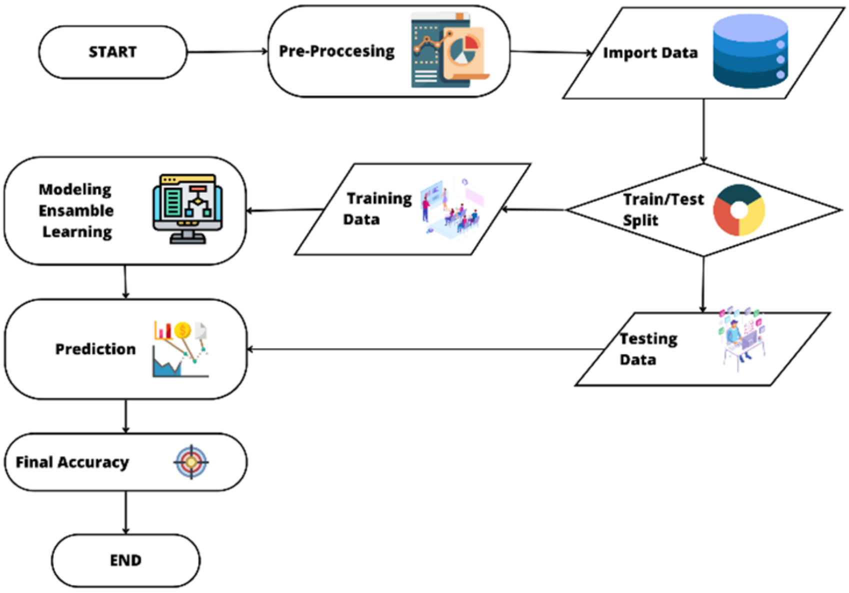


Figure 1. Flow of the proposed algorithm

2.1. Dataset Collection

The dataset used in this study is public, namely the Crude Oil WTI standard (CL=F) dataset obtained from the finance.yahoo.com website. A total of 1058 data were used in the study from January 3, 2017 to March 31, 2021 which was accessed on April 5, 2021 with West Texas Intermediate (WTI) standard size in U.S Dollars. The price used in this study is the close price because it is the price that can be a reference for predicting the open price on the next day. The data is divided into 70% training and 30% test data. The distribution of this data is based on research conducted by [12], which managed to get an accuracy rate of 99.25%. Table 1 shows the daily close price of world crude oil.

Table 1 World crude oil close price

Date	Close
1/3/2017	52.33
1/4/2017	53.26
1/5/2017	53.76
1/6/2017	53.99
1/9/2017	51.96
1/10/2017	50.82
1/11/2017	52.25
1/12/2017	53.01
1/13/2017	52.37
1/17/2017	52.48
…	…
3/31/2021	59.16

.

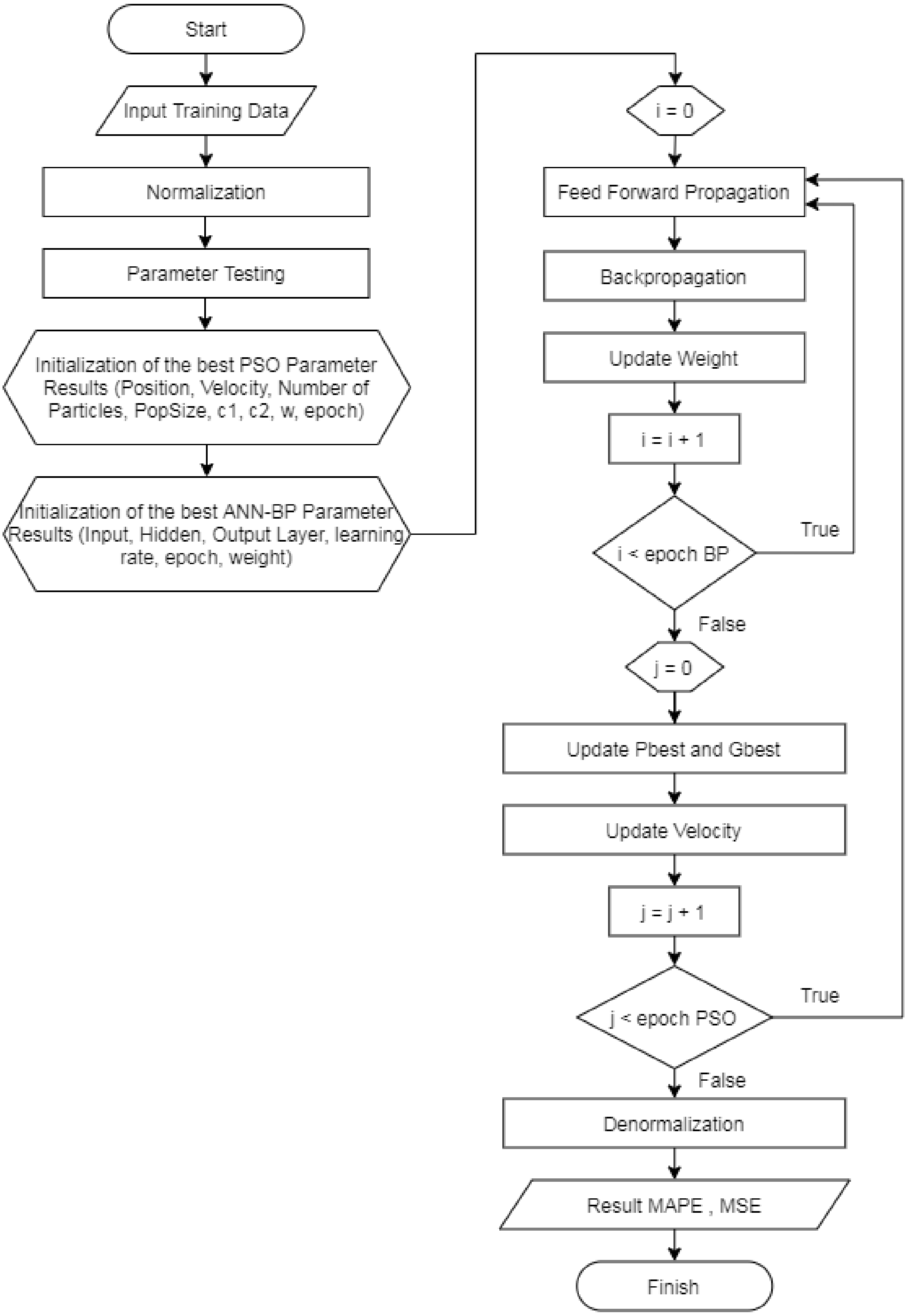


Figure 1. ANN-BP and PSO model determination flowchart

2.2. Data Normalization

In order for the method to be used to recognize data as input, it is necessary to normalize the data using a scale in the interval [0.1] in Equation 1.

$$X'_i = \frac{x_i - min}{max - min} \tag{1}$$

where,

X'_i : normalization data
x_i : data to be normalized
min : smallest data
max : biggest data

Table 2 shows the world crude oil price dataset before and after normalization in the interval [0.1].

Table 2. Dataset before and after normalization

Date	Original data	Normalized Data
1/3/2017	52.33	0.637349
1/4/2017	53.26	0.651355
1/5/2017	53.76	0.658886
1/6/2017	53.99	0.662349
1/9/2017	51.96	0.631777

2.3. Parameter Testing

At this stage, the parameters for ANN-BP and PSO parameters are tested. ANN-BP parameters tested include testing the number of input neurons and hidden neurons, the number of iterations (epochs), and the learning rate. At the same time, the PSO parameters tested include epochs and values of r1 and r2. Parameter testing was carried out using the ANN-BP training process with 70% dataset. After testing the parameter values, the best parameter values are selected through the lowest MSE and MAPE results in Table 3. The ANN-BP algorithm is shown in Figure 2 [13].

Table 3. Best parameter results PSO and ANN-BP

Parameter	Value/Amount
PSO	
Number of particles	15
Popsize	5
c_1	1
c_2	1.5
Inertia weight (w)	0.5
Epoch	16
ANN-BP	
Input layer	5
Hidden layer	3
Output layer	1
Epoch	60
Learning rate	0.2

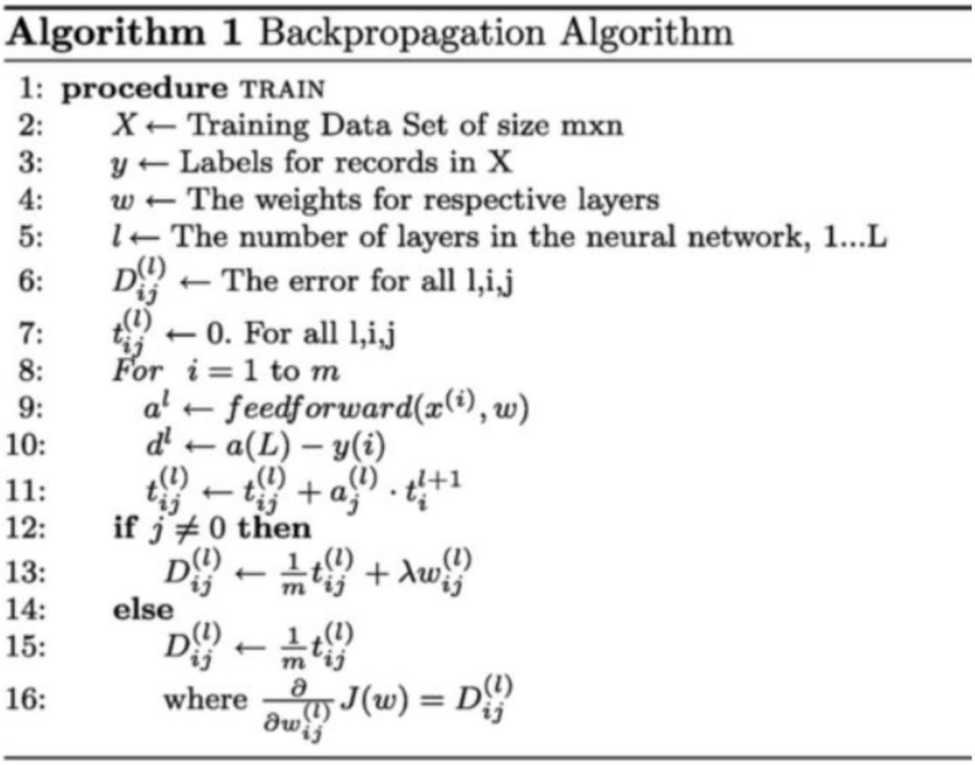
2.4. Model Determination

Determination of the model using the ANN-BP - PSO method through a training process. The dataset used is 70% of the total data. The optimization carried out by PSO in ANN-BP aims to produce the lowest error rate. PSO optimizes the ANN-BP parameters, namely weight updates so that it is expected to increase prediction accuracy. This process continues until the ANN-BP and PSO epochs have reached their limit. The execution process using this combined method takes quite a long time to adjust the number of epochs used.

.

2.5. Prediction

The prediction process is carried out using datasets as much as 30% of the total data used. The prediction process follows the ANN-BP testing stages based on the model results from the ANN-BP and PSO training processes.



```
Algorithm 1 Backpropagation Algorithm
procedure TRAIN
    $X \leftarrow$ Training Data Set of size mxn
    $y \leftarrow$ Labels for records in X
    $w \leftarrow$ The weights for respective layers
    $l \leftarrow$ The number of layers in the neural network, 1...L
    $D_{ij}^{(l)} \leftarrow$ The error for all l,i,j
    $t_{ij}^{(l)} \leftarrow 0$. For all l,i,j
    $For \;\; i = 1$ to $m$
      $a^l \leftarrow feedforward(x^{(i)}, w)$
      $d^l \leftarrow a(L) - y(i)$
      $t_{ij}^{(l)} \leftarrow t_{ij}^{(l)} + a_j^{(l)} \cdot t_i^{l+1}$
    if $j \neq 0$ then
      $D_{ij}^{(l)} \leftarrow \frac{1}{m} t_{ij}^{(l)} + \lambda w_{ij}^{(l)}$
    else
      $D_{ij}^{(l)} \leftarrow \frac{1}{m} t_{ij}^{(l)}$
        where $\frac{\partial}{\partial w_{ij}^{(l)}} J(w) = D_{ij}^{(l)}$
```

Figure 2. ANN-BP algorithm [14]

The method's success in this study is determined using indicators of predictive accuracy. These indicators are Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE). MSE is a method used to evaluate forecasting models through each error or residual squared, then summed and added to the number of observations [15]–[17]. The MSE formula can be seen in Equation 2.

$$MSE = \frac{1}{n} \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 \qquad (2)$$

where,

n : number of data points

Y_t : observed value

$\hat{Y}_t$: Predicted value

Due to its ability to be applied to various contexts, easily understood, and dependable, MAPE is regarded as the most widely used method for measuring accuracy [14], [18]. MAPE indicates how big the error is in forecasting compared to the actual value [19], [20]. The MAPE formula can be seen in Equation 3.

$$MAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|X_t - \hat{X}_t| \times 100\%}{X_t} \qquad (3)$$

where,

X_t : time series value in the t-period

$\hat{X}_t$: forecast value in the t-period

n : total number of observations

3. RESULTS AND DISCUSSIONS

The results of determining the model using the ANN-BP and ANN-BP - PSO methods, with 70% of the training data getting MSE and MAPE values, are given in Table 5. The PSO algorithm has succeeded in optimizing the weight parameter (w) in the ANN-BP training process. ANN-BP and PSO training run in each iteration. The best position is obtained, followed by updating the weight, speed, position, Pbest, and fitness to determine Gbest until the iteration is complete. The selection of parameter values determined through parameter testing has improved accuracy. The parameters obtained from the testing process were the architecture of the ANN-BP model and the PSO parameter values. The PSO parameter values comprised 15 particles, 5 popsize, an epoch value of 16, a c1 value of 1, a c2 value of 1.5, and an inertia weight value of 0.5. Meanwhile, the ANN-BP model architecture comprised 5 input layers, 3 hidden layers, 1 output layer, an epoch value of 60, and a learning rate value of 0.2.

Table 4. Results of MSE and MAPE training process

No	Metode	MSE	MAPE
1	ANN-BP	2.25938	3.03976 %
2	ANN-BP - PSO	1.96737	1.85356 %

The MSE and MAPE values generated from the training process in the search for the best parameter model using the ANN-BP and PSO methods are 1.96737 and 1.85356%, respectively. While the MSE and MAPE values using only the ANN-BP method in the training process are 2.25938 and 3.03976%, with PSO fitness results of 0.9818. These results indicate that PSO has optimized ANN-BP to get a minor error value, so the prediction model is tested in the ANN-BP training process. MSE and MAPE results from the prediction process are shown in Table 6.

Table 5. Prediction results

No	Method	MSE	MAPE
1	ANN-BP	13.86345	6.28323%
2	ANN-BP - PSO	7.15827	5.02007%

The results obtained in the prediction process are the MSE and MAPE values. The MSE and MAPE values generated by the prediction process using the ANN-BP and PSO methods are 7.15827 and 5.02007%, respectively. Meanwhile, the results of MSE and MAPE, which only used the ANN-BP method, were 13.86345 and 6.28323%. The smaller the MSE value obtained, the better the forecasting performance [21].

Although the PSO algorithm can improve the accuracy and minimize the error value in the ANN-BP method, the training process is quite time-consuming [22]. This is because each epoch in ANN-BP performs weight update calculations in each PSO epoch. Therefore, as the value of the ANN-BP and PSO epochs increases, the weight update process will also take longer.

The study using the ANN-BP-PSO model obtained better forecasting results with a high level of accuracy in predicting crude oil prices based on daily time series compared to studies that used the ARIMA method [23], Edited Nearest Neighbor (ENN) [24], Local Mean Decomposition (LMD)-ARIMA [25], and Naive [24], [25].

4. CONCLUSION

The application of the PSO algorithm in optimizing the weight parameter (w) of ANN-BP makes the prediction quality of crude oil prices increase, as evidenced by the results of MSE and MAPE ANN-BP – PSO is better than using only ANN-BP. Based on the results of the MAPE and MSE values, the testing process using the PSO algorithm in the ANN-BP method, which is 7.15827 and 5.02007%, indicates that the ANN-BP – PSO method is classified as very good and has a smaller error rate compared to using only ANN-BP method only. The prediction error value obtained decreased by 1.26316% compared to using only the ANN-BP model, which had MSE and MAPE values of 13.86345 and 6.28323% on the WTI standard Crude Oil object (CL=F).

REFERENCES

[1] E. Rizkyani, N. Aliffiyanti Iskandar, and N. Chamidah, “Klasifikasi dalam Mendeteksi Penyakit Kanker Payudara dengan Menggunakan Metode Random Forest dan Adaboost,” *Semin. Nas. Mhs. Ilmu Komput. dan Apl. Jakarta-Indonesia*, no. September, pp. 335–343, 2021.

[2] N. Sharma, K. P. Sharma, M. Mangla, and R. Rani, “Breast cancer classification using snapshot ensemble deep learning model and t-distributed stochastic neighbor embedding,” pp. 4011–4029, 2023.

[3] Z. Huang and D. Chen, “A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering

Random Forest Algorithm," *IEEE Access*, vol. 10, pp. 3284–3293, 2022, doi: 10.1109/ACCESS.2021.3139595.

[4] G. Li *et al.*, "Effective Breast Cancer Recognition Based on Fine-Grained Feature Selection," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3046309.

[5] J. Jumanto, M. F. Mardiansyah, R. Pratama, M. F. Al Hakim, and B. Rawat, "Optimization of breast cancer classification using feature selection on neural network," *J. Soft Comput. Explor.*, vol. 3, no. 2, pp. 105–110, 2022, doi: 10.52465/joscex.v3i2.78.

[6] S. Punitha, A. Amuthan, and K. S. Joseph, "Benign and malignant breast cancer segmentation using optimized region growing technique," *Futur. Comput. Informatics J.*, vol. 3, no. 2, pp. 348–358, Dec. 2018, doi: 10.1016/j.fcij.2018.10.005.

[7] A. Nugraheni, R. D. Ramadhani, A. B. Arifa, and A. Prasetiadi, "Perbandingan Performa Antara Algoritma Naive Bayes Dan K-Nearest Neighbour Pada Klasifikasi Kanker Payudara," *J. Dinda Data Sci. Inf. Technol. Data Anal.*, vol. 2, no. 1, pp. 11–20, 2022, doi: 10.20895/dinda.v2i1.391.

[8] I. Country-specific, N. Method, and M. Country-specific, "273 523 621," vol. 858, pp. 2020–2021, 2021.

[9] M. A. Muslim *et al.*, "New model combination meta-learner to improve accuracy prediction P2P lending with stacking ensemble learning," *Intell. Syst. with Appl.*, vol. 18, p. 200204, May 2023, doi: 10.1016/j.iswa.2023.200204.

[10] H. Chen, N. Wang, X. Du, K. Mei, Y. Zhou, and G. Cai, "Classification Prediction of Breast Cancer Based on Machine Learning," *Comput. Intell. Neurosci.*, vol. 2023, pp. 1–9, 2023, doi: 10.1155/2023/6530719.

[11] L. Khairunnahar, M. A. Hasib, R. H. Bin Rezanur, M. R. Islam, and M. K. Hosain, "Classification of malignant and benign tissue with logistic regression," *Informatics Med. Unlocked*, vol. 16, no. May, p. 100189, 2019, doi: 10.1016/j.imu.2019.100189.

[12] Q. Lu, Y. Li, J. Chai, and S. Wang, "Crude oil price analysis and forecasting: A perspective of 'new triangle,'" *Energy Econ.*, vol. 87, p. 104721, Mar. 2020, doi: 10.1016/j.eneco.2020.104721.

[13] Jumanto, M. F. Mardiansyah, R. N. Pratama, M. F. Al Hakim, and B. Rawat, "Optimization of breast cancer classification using feature selection on neural network," *J. Soft Comput. Explor.*, vol. 3, no. 2, pp. 105–110, Sep. 2022, doi: 10.52465/joscex.v3i2.78.

[14] S. Amar, A. Sudiarso, and M. K. Herliansyah, "The Accuracy Measurement of Stock Price Numerical Prediction," *J. Phys. Conf. Ser.*, vol. 1569, no. 3, p. 032027, Jul. 2020, doi: 10.1088/1742-6596/1569/3/032027.

[15] U. Sara, M. Akter, and M. S. Uddin, "Image Quality Assessment through FSIM, SSIM, MSE and PSNR—A Comparative Study," *J. Comput. Commun.*, vol. 07, no. 03, pp. 8–18, 2019, doi: 10.4236/jcc.2019.73002.

[16] Y. Hong, Y. Zhou, Q. Li, W. Xu, and X. Zheng, "A Deep Learning Method for Short-Term Residential Load Forecasting in Smart Grid," *IEEE Access*, vol. 8, pp. 55785–55797, 2020, doi: 10.1109/ACCESS.2020.2981817.

[17] D. Chandola, H. Gupta, V. A. Tikkiwal, and M. K. Bohra, "Multi-step ahead forecasting of global solar radiation for arid zones using deep learning," *Procedia Comput. Sci.*, vol. 167, pp. 626–635, 2020, doi: 10.1016/j.procs.2020.03.329.

[18] G. Xie, Y. Qian, and S. Wang, "Forecasting Chinese cruise tourism demand with big data: An optimized machine learning approach," *Tour. Manag.*, vol. 82, p. 104208, Feb. 2021, doi: 10.1016/j.tourman.2020.104208.

[19] M. Kumar, S. Gupta, K. Kumar, and M. Sachdeva, "SPREADING OF COVID-19 IN INDIA, ITALY, JAPAN, SPAIN, UK, US," *Digit. Gov. Res. Pract.*, vol. 1, no. 4, pp. 1–9, Oct. 2020, doi: 10.1145/3411760.

[20] S. Schreck, I. Prieur de La Comble, S. Thiem, and S. Niessen, "A Methodological Framework to support Load Forecast Error Assessment in Local Energy Markets," *IEEE Trans. Smart Grid*, vol. 11, no. 4, pp. 3212–3220, Jul. 2020, doi: 10.1109/TSG.2020.2971339.

[21] E. Windsor and W. Cao, "Improving exchange rate forecasting via a new deep multimodal fusion model," *Appl. Intell.*, vol. 52, no. 14, pp. 16701–16717, Nov. 2022, doi: 10.1007/s10489-022-03342-5.

[22] T.-A. Nguyen, H.-B. Ly, and B. T. Pham, "Backpropagation Neural Network-Based Machine Learning Model for Prediction of Soil Friction Angle," *Math. Probl. Eng.*, vol. 2020, pp. 1–11, Dec. 2020, doi: 10.1155/2020/8845768.

[23] H. S. Shambulingappa, "Crude Oil Price Forecasting Using ARIMA model," *Int. J. Adv. Sci. Inov.*, vol. 1, no. 1, pp. 1–11, 2020, doi: 10.5281/zenodo.4641697.

[24] R. Li, Y. Hu, J. Heng, and X. Chen, "A novel multiscale forecasting model for crude oil price time series," *Technol. Forecast. Soc. Change*, vol. 173, p. 121181, Dec. 2021, doi: 10.1016/j.techfore.2021.121181.

[25] J. Nasir, M. Aamir, Z. U. Haq, S. Khan, M. Y. Amin, and M. Naeem, "A New Approach for Forecasting Crude Oil Prices Based on Stochastic and Deterministic Influences of LMD Using ARIMA and LSTM Models," *IEEE Access*, vol. 11, pp. 14322–14339, 2023, doi: 10.1109/ACCESS.2023.3243232.