CLASSIFICATION OF LUNG DISEASE ON X-RAY IMAGES BASED ON GRAY LEVEL CO-OCCURRENCE MATRIX (GLCM) FEATURE EXTRACTION AND BACKPROPAGATION NEURAL NETWORK USING PYTHON GUI

Debby Mustika Rani [1], Frastica Deswardani [1], Yoza Fendriani [1*]
[1] *Department of Physics, Faculty of Science and Technology, Universitas Jambi, 36361 Indonesia*
*email: yozafendriani@unja.ac.id

ABSTRACT

This research aims to develop an automated diagnostic system for classifying *lung diseases in X-ray images based on feature extraction using the Gray Level Co-occurrence Matrix (GLCM) with a Backpropagation Artificial Neural Network employing a Python GUI. In this study, 200 lung image data were utilized, divided into four classes with 50 data points each. The four categories of image classes are normal lungs, Pneumonia, Tuberculosis, and Covid-19. The training and testing data were split in a 92:8 ratio, resulting in 184 training data and 16 testing data.* The parameters include four input layers, eight hidden layers, two output layers, alpha 0.8, 2000 iteration, and target error = 0.0001. Then, it continued with feature extraction using the GLCM to obtain texture characteristics in lung images. In the training stage, the best results were obtained in iteration 2000 with a Mean Squared Error of 0.005% and a calculated time of 167.319 seconds. At the testing stage, a reasonably high accuracy was obtained, 93.75%, with a calculated time of 0.014 seconds. This result indicates that the method can prove lung images.

Keywords: *Backpropagation Neural Network; Feature Extraction; GLCM; Image Classification; Phyton GUI*

INTRODUCTION

The lung is an essential organ for human respiration. Lung diseases affecting these organs can significantly impact an individual's health and well-being. Detecting abnormality or disease of the lungs is generally done clinically by a doctor. Lung disease can identified with the image of lungs obtained from X-ray images. Lung imaging can also be obtained using a CT-Scan (Computed Tomography Scan) and MRI (Magnetic Resonance Imaging). Possible lung problems are Tuberculosis (TB), Pneumonia, Coronavirus Disease-19 (Covid-19), etc.

Accurate and timely diagnosis of lung diseases is crucial for effective treatment and management. Traditionally, pulmonary diagnosis relies on a combination of clinical presentation, imaging studies, and invasive procedures such as histological biopsies. In recent years, research has explored the application of computer-aided diagnosis systems in response to the potential risks associated with histological biopsies. These systems aim to augment radiologist's diagnostic capabilities by providing objective analysis and improving diagnostic accuracy (Anthimopoulos et al., 2016).

Backpropagation Neural Networks (BPNN) have emerged as a powerful tool for pattern recognition and classification tasks, making them well-suited for analyzing medical images and identifying abnormalities associated with lung diseases. The BPNN model is a multilayer forward neural network that can realize the random nonlinear mapping of corresponding input and output as well as the autonomous learning, so it has emerged in the processing and solution of nonlinear issues (Wu et al., 2019; Alcantara et al., 2019). BPNNs are a type of artificial neural network inspired by the structure and function of the human brain. They consist of interconnected layers of neurons that process and transmit information. The training process involves feeding the network with labeled data, allowing it to learn the underlying patterns and relationships between features. Once trained, the BPNN can classify new data points with high accuracy. (Wang et al. 2015).

To further enhance the diagnostic capabilities of BPNN, feature extraction techniques are employed to extract relevant and discriminative information from medical images. Gray Level Co-occurrence Matrix (GLCM) techniques analyze texture or extract images on a specific pattern. The GLCM is a statistical technique for extracting texture features from grayscale images. It quantifies the spatial distribution of gray level intensities within an image by calculating the probabilities of co-occurring gray level pairs at a specified distance (d) and angle (θ). (Kshirsagar et al, 2020; Asery et al, 2016). By analyzing the distribution of pixel pairs with similar intensities, GLCM provides valuable insights into the texture and structure of the tissue, making it suitable for identifying abnormalities in lung images.

Several studies have explored the application of artificial intelligence and machine learning for pulmonary disease diagnosis using medical images.

Depinta and Abdullah (2017) employed a backpropagation neural network (BPNN) model for tuberculosis detection based on chest x-ray data. Their model achieved an accuracy of 79.41% using 34 test data and 30 training data. Hamid (2019) developed a convolutional neural network (CNN) model for classifying tuberculosis and pneumonia. The model attained an accuracy of 85.96%. Putra and Yuhandri (2021) investigated the identification of COVID-19 patients using chest X-rays and a BPNN algorithm. Their model achieved an accuracy of 73% using GLCM features, including correlation, energy, contrast, and homogeneity.

The studies, as mentioned earlier, demonstrate the potential of artificial intelligence and machine learning techniques for pulmonary disease diagnosis. However, there is still room for improvement in accuracy and generalizability. Building upon prior research, this study investigates the classification of four lung conditions: pneumonia, tuberculosis, COVID-19, and normal lung. This expanded dataset allows for a more comprehensive comparison during the classification of lung diseases using chest X-rays. Gray Level Co-occurrence Matrix (GLCM) feature extraction, encompassing energy, contrast, entropy, and homogeneity features, will be employed with a Backpropagation neural network (BPNN) classifier. This study aims to explore further the application of artificial intelligence and machine learning for pulmonary disease diagnosis, mainly focusing on enhancing accuracy.

METHOD

The study utilized a dataset of grayscale JPEG images from the Kaggle website (https://www.kaggle.com) and chest X-ray images from Bhayangkara Hospital Jambi, Indonesia. The dataset comprised 200 images, with 50 images per class. The training and testing sets were divided in a 92:8 ratio, resulting in 184 training images and 16 testing images. The images were categorized into four classes: Pneumonia, Tuberculosis, COVID-19, and Normal lung. The research employed the following tools and software:

- Programming Language: Python
- GUI Design Tool: Qt Designer
- Integrated Development Environment (IDE): Visual Studio Code
- Data: Chest X-ray images
- Libraries: Pandas, Numpy, Seaborn, Statsmodels, Matplotlib, Scikit Learn
- Hardware: Computer

GUI Design using Qt Designer

The graphical user interface (GUI) was designed using Qt Designer. The GUI design involved creating parameters for the input layer, hidden layer, output layer, learning rate, error tolerance, and iterations. These parameters were used during the training process and for designing the testing process for lung disease classification. The testing process allowed for the evaluation of the classification accuracy and error difference.

Development Backpropagation Neural Network Class

Implementing a Backpropagation neural network for lung disease classification and prediction involved the creation of a Backpropagation neural network class. This class encapsulates the functions that represent each stage of the Backpropagation neural network algorithm.

Image Processing

The Image Processing class was developed to handle image processing operations, including data reading, image display, and feature extraction. The data reading function reads the chest X-ray images from the training and testing datasets. The images are stored in a format compatible with the neural network model. Image display allows users to visualize the images and assess their quality before proceeding with feature extraction and classification.

Executing the Backpropagation Neural Network GUI

The Backpropagation neural network GUI is executed to classify lung diseases. Upon launching the GUI, the application reads the training data, including chest X-ray images of COVID-19, Pneumonia, Tuberculosis, and Normal cases. The neural network's initial weights, V and W, are initialized, and the training process commences. The training process involves feeding the neural network with labelled chest X-ray images and adjusting the weights to minimize the classification error.

Feature Extraction

Feature extraction is a crucial step in image analysis and classification tasks. It involves extracting meaningful characteristics from the raw image data to represent the underlying patterns and information. These extracted features are then used as input to machine learning algorithms for further analysis and classification. Extraction features used *Gray Level Co-occurrence Matrix* (GLCM) method.

Data Training Process

The training process is an integral step in Backpropagation neural network learning, where the network learns to recognize patterns from the training data. During training, the network's weights are adjusted to minimize the error between the predicted outputs and the actual target outputs. The weight initialization process must determine the parameters of the neurons that will be used, namely the input layer, hidden layer and output layer. Then, the training process results are V weights, W weights, trained biases, MSE, and computing time during training. The trained weights are the representation and knowledge possessed by the artificial neural network.

Data Testing Process

The data testing process is a crucial step in evaluating the performance of a trained Backpropagation neural network. It involves testing the network's ability to classify or predict unseen data, thereby assessing its generalization capability. The testing data comprises a separate set of images that were not used during the training process. This ensures that the network's performance is evaluated on data it has never encountered before. The testing data should be representative of the real-world data the network will be used to classify.

Classification / Prediction

The extracted features from the testing data are fed into the trained neural network, and the network generates predictions for the corresponding lung disease categories. These predictions are compared to the actual labels in the testing data to evaluate the network's performance. This evaluation determines whether the network can accurately classify unseen chest X-ray images according to the lung disease categories.

RESULTS AND DISCUSSION

1. Training Process

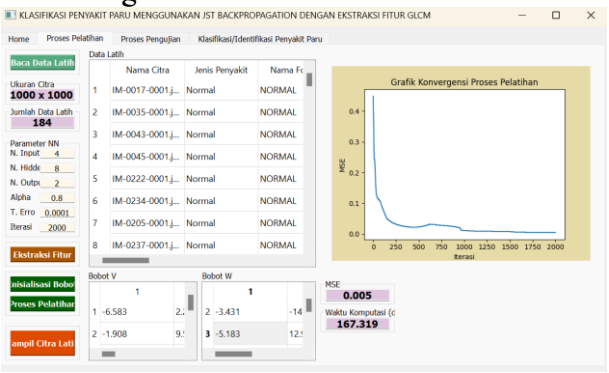

Figure 1. Training Process GUI Display

Figure 1 shows the training process, involving reading the training data after creating lung disease classification classes and then running the Backpropagation neural network GUI application to classify lung diseases. The training process was repeated four times with different iterations to find the smallest MSE. When the first training was performed at 500 iterations, the MSE obtained was 0.023, indicating an error of 2.3% with a computation time of 41.993 seconds. The second training was then performed at 1000 iterations, resulting in an MSE of 0.011, indicating an error of 1.1% with a computation time of 83.804 seconds. The third training was performed at 1500 iterations and resulted in an MSE of 0.008, indicating an error of 0.8% with a computation time of 125.829 seconds. The fourth training was performed at 2000 iterations, resulting in an MSE of 0.005, indicating an error of 0.5% with a computation time of 167.319 seconds. The error value obtained in the fourth training was quite low compared to the first, second, and third trainings. Therefore, there will be no obstacles that will affect the testing process.

Based on the training results, it can be concluded that the training with 2000 iterations achieved the smallest MSE, indicating that the best MSE was found among the four trainings performed. However, the computation time obtained in the fourth training was significantly higher than in the first, second, and third training. This indicates that the higher the iteration, the more time it takes to complete the training process.

Furthermore, the alpha value (learning rate) is one of the parameters in the Backpropagation neural network algorithm, where its value ranges from 0 to 1 (Wadi, 2021). The larger the learning rate value, the faster the training process will be. Therefore, the value can be changed to obtain the best training results.

Tables 1 and 2 present the values of weight initialization using bias. Bias is employed to accelerate the learning rate (Gonzalez, 2009). Consequently, the number of V weights and W weights will change due to the addition of 1 to the number of input layer neurons in the V weights and the number of hidden layer neurons in the W weights. In weight initialization, the number of neurons in the input layer is 4+1=5 and the hidden layer is 8, so the number of V weights connecting the two layers becomes 40 (5x8). For the W weights for the hidden layer, the number of neurons is 8+1=9 and the output layer is 2, so the number of W weights connecting the two layers becomes 18 (9x2).

Table 1 . Initialization V weight

i \ j	1	2	3	4	5	6	7	8
1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
2	0.058	0.762	0.833	0.039	0.019	0.238	0.797	0.084
3	0.929	0.974	0.785	0.684	0.903	0.48	0.42	0.278
4	0.893	0.504	0.045	0.22	0.851	0.945	0.144	0.445
5	0.897	0.913	0.819	0.666	0.725	0.205	0.797	0.876

Table 2. Initialization W weight

i \ j	1	2
1	0.1	0.1
2	0.94	0.356
3	0.106	0.917
4	0.662	0.163
5	0.7	0.129
6	0.825	0.942
7	0.849	0.925
8	0.118	0.133
9	0.865	0.288

2. Data Testing Process

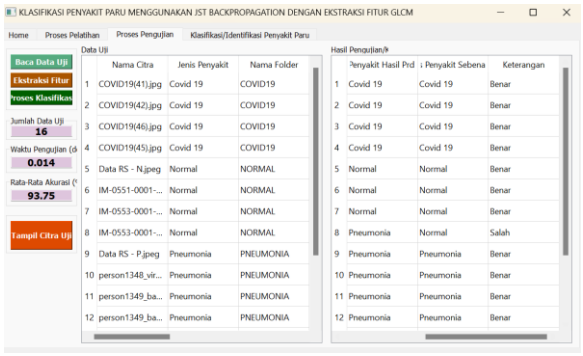

Figure 2. GUI Display of Testing Process

Figure 2 shows the testing process, involving reading the test data. During the testing process, a high accuracy of 93.75% was achieved. This indicates that the low MSE values obtained during training positively influenced the accuracy of the testing process. Another factor contributing to the high accuracy is the abundance of data. Generally, more data leads to better training outcomes, as evidenced by lower MSE values.

The trend of improved results with increasing iteration count also holds true for the Backpropagation neural network. As the number of iterations increases, the network has more opportunities to refine its weights and minimize the error, leading to enhanced classification performance. However, this comes at the expense of increased computation time. The testing process, which involves evaluating the trained network's performance on unseen data, requires more computational resources as the iteration count increases. This is because the network with a higher number of iterations has a more complex weight structure that needs to be evaluated for each testing instance. In the case of the Backpropagation neural network used in this study, the computation time for testing was 0.014 seconds, indicating a relatively efficient training process.

Table 3. Test Image Data Source

Name	Amount of data	Information
Covid-19	4	4 Kaggle datasets
Normal	4	3 Kaggle datasets 1 Hospital data
Pneumonia	4	3 Kaggle datasets 1 Hospital data
Tuberculosis	4	3 Kaggle datasets 1 Hospital data

Table 3 presents the source of the test data employed for the evaluation process. For each category of normal, pneumonia, and tuberculosis

chest X-ray images, three data were obtained from the Kaggle dataset and one data from the hospital dataset. The hospital chest X-ray images served as a comparison benchmark. In the case of COVID-19 chest X-ray images, four data points were retrieved from the Kaggle dataset due to the unavailability of COVID-19 hospital data at the time of the study.

Table 4. Feature Extraction Value (GLCM)

Name	Code	Energy	Contrast	Entropy	Homogeneity
Covid-19 (1)	00	0.012	8,767	5,768	0.754
Covid-19 (2)	00	0.009	8,615	5,997	0.724
Covid-19 (3)	00	0.002	7,499	6.92	0.622
Covid-19 (4)	00	0.004	11,944	7,175	0.631
Normal Hospital Data	01	0.014	20,118	7,327	0.402
Normal (1)	01	0.006	19.102	7,616	0.361
Normal (2)	01	0.007	24,064	7,761	0.331
Normal (3)	01	0.012	17.29	7,509	0.38
Pneumonia Hospital Data	10	0.002	17,405	7,096	0.508
Pneumonia (1)	10	0.005	16,556	7,372	0.464
Pneumonia (2)	10	0.003	16,525	6,742	0.528
Pneumonia (3)	10	0.002	12,854	7,149	0.489
Tuberculosis Hospital data	11	0.006	3,723	6,676	0.691
Tuberculosis (1)	11	0.019	3,828	6,443	0.691
Tuberculosis (2)	11	0.003	1,388	6,302	0.758
Tuberculosis (3)	11	0.002	3,296	6,701	0.704

Table 4 presents the feature extraction results for the test data images. The first column represents the energy value, where energy represents a measure of uniformity in an image (Hadi & Rachmawanto, 2022). Images with fewer grayscale levels will have higher energy compared to images with many grayscale levels (Maulida et al., 2022). From the table above, the image with the fewest grayscale levels is the Tuberculosis (1) with an energy value of 0.019. There are 4 images with many grayscale levels, including the COVID-19 (3) image, the hospital Pneumonia image, the Pneumonia (3) image, and the Tuberculosis (3) image, with results of 0.002. Based on the diverse energy values of the feature extraction for each image, recognition using this method is possible.

The second column of Table 4 displays the contrast values, a measure of grayscale intensity variation within an image. If a pixel and its neighboring pixels have close intensity (grayscale) values, so the texture contrast is very low (valued at 0). Conversely, if a pixel and its neighboring pixels have distant intensity (grayscale) values, so the texture contrast is high (Listia, 2014). The Normal (2) image has the highest contrast value compared to other lung images 24.064. Based on the diverse contrast values of the feature extraction for each image, recognition using this method is possible.

The third column of Table 4 presents the entropy values, which measure image complexity. Entropy values tend to be higher for images with greater non-uniformity. Conversely, they exhibit an inverse relationship with energy. The table shows that the Normal (2) image holds the highest entropy value (7.761) compared to other lung images. This diversity in entropy values across the images suggests the potential of this feature for image recognition using the proposed method.

The final column in Table 4 presents the homogeneity values, representing a uniformity measure within the co-occurrence matrix. High homogeneity values indicate that all pixels have similar values. As the table shows, a homogeneity value approaching 1 implies greater uniformity in intensity or grayscale levels. Conversely, low homogeneity values suggest lower image similarity (Maulida et al., 2022). The Tuberculosis (2) image exhibits a high homogeneity value of 0.758, indicating that its intensity or grayscale levels are more uniform. The diverse homogeneity values across the images suggest the potential of this feature for image recognition using the proposed method.

Furthermore, the study reveals that the contrast values of individual images are generally higher compared to the values of other features, which exhibit less variation between images. This suggests that contrast significantly influences the testing and training processes. Following feature extraction, the classification process is conducted to determine and verify whether the classification results for the lung images align with the actual diseases.

Table 5. Display of Classification Process Results

Image Name	Code	Disease Prediction Results	Actual Disease	Information
Covid-19 (1)	00	Covid	Covid	Correct
Covid-19 (2)	00	Covid	Covid	Correct
Covid-19 (3)	00	Covid	Covid	Correct
Covid-19 (4)	00	Covid	Covid	Correct
Normal Hospital Data	01	Normal	Normal	Correct
Normal (1)	01	Normal	Normal	Correct
Normal (2)	01	Normal	Normal	Correct
Normal (3)	01	Pneumonia	Normal	Wrong
Pneumonia Hospital Data	10	Pneumonia	Pneumonia	Correct
Pneumonia (1)	10	Pneumonia	Pneumonia	Correct
Pneumonia (2)	10	Pneumonia	Pneumonia	Correct
Pneumonia (3)	10	Pneumonia	Pneumonia	Correct
Tuberculosis Hospital Data	11	Tuberculosis	Tuberculosis	Correct
Tuberculosis (1)	11	Tuberculosis	Tuberculosis	Correct
Tuberculosis (2)	11	Tuberculosis	Tuberculosis	Correct
Tuberculosis (3)	11	Tuberculosis	Tuberculosis	Correct

Table 5 shows results from the image classification process lungs with 16 data of testing images. The classification results revealed one misclassified lung image out of 16, namely the Normal (3) lung image. This image was incorrectly classified as pneumonia, while the actual data indicated a normal lung image. This indicates that the classification accuracy is not yet 100% perfect, as one lung image does not align with the actual disease type. In contrast, the classification process applied to the hospital images yielded accurate results consistent with the actual diseases. Therefore, it can be concluded that the classification process using the Backpropagation neural network can effectively identify diseases from lung images.

3. Classification Data Results

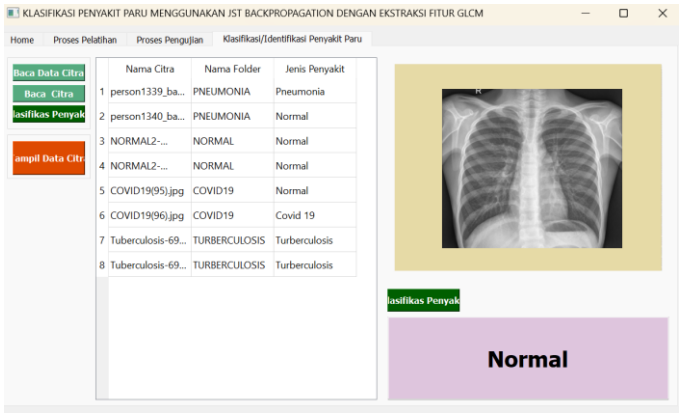

Figure 3. GUI display of Image Classification

Figure 3 illustrates the classification or identification process for lung diseases. The classification process is employed to identify the disease of an image with an unknown disease type, following a successful testing phase with high accuracy. In this classification process, eight image data are utilized. After the disease classification is performed, the results indicate that six out of eight data match the actual disease. However, the image data for the second and fifth cases does not align with the actual disease, being identified as normal lung images. This highlights that the less-than-perfect accuracy in the testing phase impacts the classification of lung diseases.

CONCLUSIONS

The research revealed that the extracted contrast values significantly influence the training process. Therefore, organizing the contrast values from the highest to the lowest is crucial. If the contrast values are not arranged or are randomly distributed, the resulting MSE (Mean Squared Error) will be considerably higher. Additionally, the MSE is influenced by the number of training data points. A larger dataset generally leads to better training outcomes, indicated by a lower MSE. Similarly, increasing the number of iterations tends to improve the MSE. During the testing process, one image from the Kaggle dataset was misclassified, indicating that the classification accuracy is not yet 100%, with an achieved accuracy of 93.75%. In contrast, the classification process for hospital images accurately identified the corresponding diseases. This suggests that the Backpropagation Neural Network classification method can effectively diagnose diseases from lung images. Moreover, the system successfully extracts features, obtaining diverse values for energy, contrast, entropy, and homogeneity in the Gray Level Co-occurrence Matrix (GLCM) for each image, demonstrating the feasibility of image recognition using this method.

BIBLIOGRAPHY

Alcantara, J. H., & Chen, J. S. (2019). Neural networks based on three classes of NCP-functions for solving nonlinear complementarity problems. *Neurocomputing*, *359*, 102-113.

Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., & Mougiakakou, S. (2016). Lung Pattern Classification for Interstitial Lung Diseases Using a Deep Convolutional Neural Network. IEEE transactions on medical imaging, 35(5), 1207–1216.

Asery, R., Sunkaria, R. K., Sharma, L. D., & Kumar, A. (2016, June). Fog detection using GLCM-based features and SVM. In 2016 Conference on Advances in Signal Processing (CASP) (pp. 72-76).

Depinta, L., dan Abdullah, Z. 2017. Implementasi Jaringan Syaraf Tiruan Backpropagation untuk Deteksi Penyakit Tuberculosis (TB) Paru dari Citra Rontgen. *Jurnal Fisika Unand*. 6(1).

Gonzalez, R. C. (2009). Digital image processing fourth edition. New York: Pearson Publisher.

Hadi, H. P., & Rachmawanto, E. H. (2022). Ekstraksi Fitur Warna Dan Glcm Pada Algoritma Knn Untuk Klasifikasi Kematangan Rambutan. Jurnal Informatika Polinema, 8(3), 63-68.

Hamid, A. (2019). Klasifikasi Penyakit Tuberculosis dan Pneumonia pada Paru-Paru Manusia Berdasarkan Citra Chest X-Ray Menggunakan Convolutional Neural Network (Bachelor's thesis, Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta).

Kshirsagar, P. R., Yadav, A. D., Joshi, K. A., Chippalkatti, P., & Nerkar, R. Y. (2020). Classification and detection of brain tumor by using GLCM texture feature and ANFIS. J. Res. Image Signal Processing, 5, 15-31.

Listia, R., dan Harjoko, A. 2014. Klasifikasi Massa pada Citra Mammogram Berdasarkan Gray Level Co-occurrence Matrix (GLCM). IJCCS (Indonesian Journal of Computing and Cybernetics Systems). 8(1).

Maulida, A., Nurhidayah, N., Fendriani, Y., & Haryono, H. (2022). Segmentasi Citra Mammogram Untuk Deteksi Dini Kanker Payudara Dengan Menggunakan Metode Otsu Thresholding. Jurnal Fisika Unand, 11(2), 180-186.

Putra, H. R. W., & Yuhandri, Y. (2021). Identifikasi Penderita COVID-19 Berdasarkan Chest X-ray Menggunakan Algoritma Jaringan Syaraf Tiruan Backpropagation. Jurnal Sistim Informasi dan Teknologi, 197-202.

Wadi, H. 2021. Klasifikasi Citra dengan Jaringan Syaraf Tiruan Backpropagation menggunakan Python GUI. Nusa Tenggara : Turida Publisher.

Wang, L., Zeng, Y., & Chen, T. (2015). Back propagation neural network with adaptive differential evolution algorithm for time series forecasting. *Expert Systems with Applications*, *42*(2), 855-863.

Wu, L., Yang, Y., Maheshwari, M., & Li, N. (2019). Parameter optimization for FPSO design using an improved FOA and IFOA-BP neural network. Ocean Engineering, 175, 50–61.