Institute of Research and Publication Indonesia
IJIRSE: Indonesian Journal of Informatic Research and Software Engineering
Journal Homepage: https://journal.irpi.or.id/index.php/ijirse
Vol. 2. No.1. Maret 2022, pp: 67 - 72
E-ISSN: 2775-5754 | P-ISSN: 2797-2712

The Effect of Data Normalization on the Performance of the Classification Results of the Backpropagation Algorithm

Pengaruh Normalisasi Data Terhadap Performa Hasil Klasifikasi Algoritma Backpropagation

Inggih Permana[1*], Febi Nur Salisah[2]

[1,2]Sistem Informasi, Fakultas Sains dan Teknologi,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia

E-Mail: [1]inggihpermana@uin-suska.ac.id, [2]febinursalisah@uin-suska.ac.id

Makalah: Diterima 21 Januari 2022; Diperbaiki 28 Maret 2022; Disetujui 28 Maret 2022
Corresponding Author: Inggih Permana

Abstrak

Keberhasilan Algoritma Backpropagation (BP) tergantung pada kualitas data. Sehingga, normalisasi data merupakan proses yang penting. Akan tetapi, beberapa penelitian juga ada yang tidak menggunakan normalisasi data. Oleh sebab itu, penelitian ini mengukur pengaruh normalisasi data terhadap performa hasil klasifikasi Algoritma Backpropagation. Agar diketahui apakah normalisasi benar-benar bisa meningkatkan performa hasil klasifikasi pada Algoritma BP. Penelitian ini menggunakan tiga metode normalisasi data, yaitu: MinMax Normalization; MaxAbs Normalization; dan Z-Score Normalization. Berdasarkan hasil percobaan didapat bahwa jika data yang digunakan terdapat perbedaan rentang nilai antar atribut yang tidak berbeda jauh, maka BP tanpa normalisasi data bisa menjadi pilihan terbaik. Akan tetapi jika pada data terdapat atribut yang memiliki perbedaan rentang nilai yang jauh dari atribut lainnya, maka menggunakan normalisasi data bisa menjadi pilihan terbaik. Berdasarkan hasil percobaan juga didapat bahwa Z-Score Normalization merupakan metode normalisasi terbaik.

Keyword: backpropagation, normalisasi data, performa

Abstract

The success of the Backpropagation Algorithm (BP) depends on the quality of the data. Thus, data normalization is an important process. However, there are also some studies that do not use data normalization. Therefore, this study measures the effect of data normalization on the performance of the classification results of the Backpropagation Algorithm. In order to know whether normalization can really improve the performance of the classification results on the BP Algorithm. This study uses three methods of data normalization, namely: MinMax Normalization; MaxAbs Normalization; and Z-Score Normalization. Based on the experimental results, it is found that if there are differences in the range of values between attributes that do not differ much, then BP without data normalization can be the best choice. However, if there are attributes in the data that have different value ranges that are far from other attributes, then using data normalization can be the best choice. Based on the experimental results, it is also found that Z-Score Normalization is the best normalization method.

Keyword: backpropagation, data normalization, performance

1. Pendahuluan

Keberhasilan sebuah Metode *Machine Learning* (ML) tergantung pada kualitas data [1]. Oleh sebab itu, fase pra proses data merupakan proses yang sangat penting untuk meningkatkan performa ML [2]. Normalisasi data merupakan salah satu proses yang dilakukan pada fase pra proses data. Pada normalisasi dilakukan penskalaan kembali nilai-nilai sehingga dapat membuat pemrosesan lebih mudah [3]. Selain itu, normalisasi data tidak menyebabkan peningkatan besar dalam beban kerja memori dan kebutuhan daya untuk pemrosesan [4].

Artificial Neural Network (ANN) adalah salah satu metode yang sering digunakan dalam ML. Metode ini terinspirasi dari cara kerja jaringan syaraf manusia. Algoritma *Backpropagation* (BP) adalah algoritma pembelajaran yang sering digunakan dalam metode ANN. Algoritma BP adalah algortima pembelajaran ANN yang paling populer [5] yang diusulkan oleh Rumelhart, Hinton dan Williams [6]. BP menggunakan konsep gradient descent untuk melakukan pembelajaran.

BP memerlukan normalisasi data yang sesuai untuk mendapatkan performa terbaik [7]. Akan tetapi, ada juga beberapa penelitian yang tidak melakukan normalisasi data untuk proses pelatihan ML [8]. Oleh sebab itu penelitian ini mengukur pengaruh normalisasi data terhadap performa hasil klasifikasi dari Algoritma BP. Penelitian ini membandingkan tiga metode normalisasi data yang sering digunakan, yaitu: MinMax *Normalization*; Z-Score *Normalization*; dan MaxAbs *Normalization*. Hal ini dilakukan agar bisa mengetahui metode normalisasi yang cocok untuk BP. Selain itu, performa Algoritma BP dengan data dinormalisasi juga dibandingkan dengan data yang tidak dinormalisasi. Hal ini dilakukan agar bisa melihat apakah normalisasi data benar-benar membuat performa BP menjadi lebih baik atau tidak. Penelitian ini menggunakan tiga dataset, yaitu Dataset Iris, Dataset *Wine*, dan Dataset *Breast Cancer*. Hal ini dilakukan agar bisa melihat apakah pengaruh normalisasi berlaku di banyak kondisi data atau hanya berlaku di data tertentu saja.

Penelitian terdahulu sudah ada yang mengukur pengaruh normalisasi data terhadap performa ML. Akan tetapi penelitian-penelitian tersebut menggunakan data kurang dari tiga dataset [2, 9, 10, 11, 12, 13, 14], sehingga hasil penelitian-penelitian tersebut tidak bisa memperlihatkan apakah hasil tersebut berlaku untuk banyak kondisi data atau hanya terbatas di data tersebut saja. Ada juga penelitian-penelitian yang tidak membandingkan performa ML antara data dinormalisasi dan data tidak dinormalisasi [2, 7, 9, 11, 15], sehingga penelitian-penelitian tersebut tidak bisa memperlihatkan apakah data yang dinormalisasi benar-benar menghasilkan performa ML yang lebih baik atau tidak.

Paper ini terdiri dari 5 bab. Bab 2 membahas tentang material dan metode yang digunakan. Bab 4 membahas hasil percobaan. Sedangkan bab terakhir berisi kesimpulan dari penelitian ini.

2. Material dan Metode

Bab ini akan dijelaskan tentang metode-metode dan material-material yang digunakan pada penelitian ini, seperti dataset-dataset yang digunakan, metode normalisasi yang digunakan, penerapan Algoritma BP, pengukuran performa, dan tool yang digunakan.

2.1 Dataset-Dataset

Sub Bab ini berisi tentang deskripsi singkat dari dataset-dataset yang digunakan pada penelitian ini.

2.1.1. Dataset Iris

Dataset Iris berisi data tanaman-tanaman iris. Dataset ini awalnya digunakan pada penelitian pengukuran pada permasalahan taksonomi yang dilakukan oleh Fisher [16]. Dataset ini terdiri dari 150 baris data. Dataset ini terdiri dari 3 kelas yaitu: (1) iris setosa; (2) iris versicolor; dan (3) iris virginica. Masing-masing kelas tersebut terdiri dari 50 baris data. Terdapat 4 atribut pada dataset ini, yaitu: (1) sepal *length* (cm); (2) sepal *width* (cm), (3) petal *length* (cm), dan (4) petal *width* (cm).

2.1.2. Dataset *Wine*

Dataset *Wine* merupakan data hasil analisis kimia dari anggur yang ditanam di wilayah yang sama di Italia oleh tiga pembudidaya berbeda. Pemilik asli dari data ini adalah Forina [17]. Dataset ini terdiri dari 178 baris data. Dataset ini terdiri dari 3 kelas yaitu: (1) *class*_0 (59); (2) *class*_1 (71); dan (3) *class*_2 (48). Terdapat 13 atribut pada dataset ini, yaitu: (1) *alcohol*; (2) *malic acid*; (3) *ash*; (4) *alcalinity of ash*; (5) *magnesium*; (6) total *phenols*; (7) *flavanoids*; (8) *nonflavanoid phenols*; (9) *proanthocyanins*; (10) *color intensity*; (11) *hue*; (12) OD280/OD315 *of diluted wines*; dan (13) *proline*.

2.1.3. Dataset *Breast Cancer*

Dataset *Breast Cancer* merupakan kumpulan data kanker payudara. Dataset ini dibuat oleh Wolberg dkk. [18, 19]. Dataset ini terdiri dari 569 baris data. Dataset ini terdiri dari 2 kelas, yaitu WDBC-Malignant dan WDBC-Benign. Terdapat 30 atribut pada dataset ini, yaitu: (1) *radius* (*mean*); (2) *texture* (*mean*); (3) *perimeter* (*mean*); (4) *area* (*mean*); (5) *smoothness* (*mean*); (6) *compactness* (*mean*); (7) *concavity* (*mean*); (8) *concave points* (*mean*); (9) symmetry (mean); (10) fractal dimension (mean); (11) *radius* (*standard error*); (12) *texture* (*standard error*); (13) *perimeter* (*standard error*); (14) *area* (*standard error*); (15) *smoothness* (*standard error*); (16) *compactness* (*standard error*); (17) *concavity* (*standard error*); (18) *concave points* (*standard error*); (19) *symmetry* (*standard error*); (20) *fractal dimension* (*standard error*); (21) *radius* (*worst*); (22)

texture (*worst*); (23) *perimeter* (*worst*); (24) *area* (*worst*); (25) *smoothness* (*worst*); (26) *compactness* (*worst*); (27) *concavity* (*worst*); (28) *concave points* (*worst*); (29) *symmetry* (*worst*); dan (30) *fractal dimension* (*worst*).

2.2. Metode-Metode Normalisasi

Seperti yang telah disinggung pada Bab 1, penelitian ini fokus pada tiga metode normalisasi, yaitu: MaxAbs *Normalization*, MinMax *Normalization*, dan Z-*Score Normalization*. Metode-metode tersebut akan dijelaskan pada tiga sub bab berikut ini.

2.2.1. MinMax *Normalization*

Metode MinMax *Normalization* adalah metode normalisasi yang merubah rentang nilai data menjadi antara 0 dan 1. Persamaan untuk menghitung MinMax *Normalization* dapat dilihat pada Persamaan 1 [20].

$$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Keterangan:

1. x_i = nilai tertentu yang akan dinormalisasi
2. x' = nilai hasil normalisasi
3. min(x) = nilai minimal dari sebuah atribut
4. max(x) = nilai maksimal dari sebuah atribut

2.2.2. Z-*Score Normalization*

Metode Z-*Score Normalization* menggunakan *mean* dan standar deviasi pada setiap atribut fitur untuk merubah skala nilai dari data [20]. Metode normalisasi ini dapat mengurangi efek outlier. Persamaan untuk menghitung Z-*Score Normalization* dapat dilihat pada Persamaan 2 [20].

$$x' = \frac{x_i - \text{mean}(x)}{\text{std}(x)} \quad (2)$$

Keterangan:

1. x_i = nilai tertentu yang akan dinormalisasi
2. x' = nilai hasil normalisasi
3. mean(x) = nilai rata-rata dari sebuah atribut
4. std(x) = nilai standar deviasi dari sebuah atribut

2.2.3. MaxAbs *Normalization*

Metode MaxAbs *Normalization* adalah metode normalisasi data yang membagi semua nilai dengan nilai absolut dari nilai maksimum. Hal ini merubah nilai maksimum menjadi 1. Metode ini tidak mengubah sparitas data karena metode ini tidak memusatkan data. Persamaan untuk menghitung MaxAbs *Normalization* dapat dilihat pada Persamaan 3 [20].

$$x' = \frac{x_i}{|\max(x)|} \quad (3)$$

Keterangan:

1. x_i = nilai tertentu yang akan dinormalisasi
2. x' = nilai hasil normalisasi
3. max(x) = nilai maksimal dari sebuah atribut

2.3. Penerapan Algoritma BP

Nilai-nilai pada parameter-parameter Algoritma BP yang digunakan pada penelitian ini sebagai berikut:

1. Nilai learning rate inisial: 0.01
2. Jumlah Iterasi maksimal: 200
3. Jumlah hidden layer dan node-nya: 1 hidden layer dengan 100 node
4. Fungsi aktifasi: Rectified Linear Activation Unit (ReLU)

Pengukuran performa hasil klasifikasi Algoritma BP dilakukan dengan menggunakan akurasi. Sedangkan untuk validasi, penelitian ini menggunakan K-*Fold Cross Validation* dengan nilai K = 5.

2.5. *Tool*

Penelitian ini menggunakan Bahasa Pemrograman Python. Library *machine learning* yang digunakan adalah Scikit-Learn. Pada *library* ini sudah terdapat dataset-dataset, algoritma, serta teknik validasi yang dilakukan pada penelitian ini. Lingkungan pengembangan yang digunakan adalah Google Colab

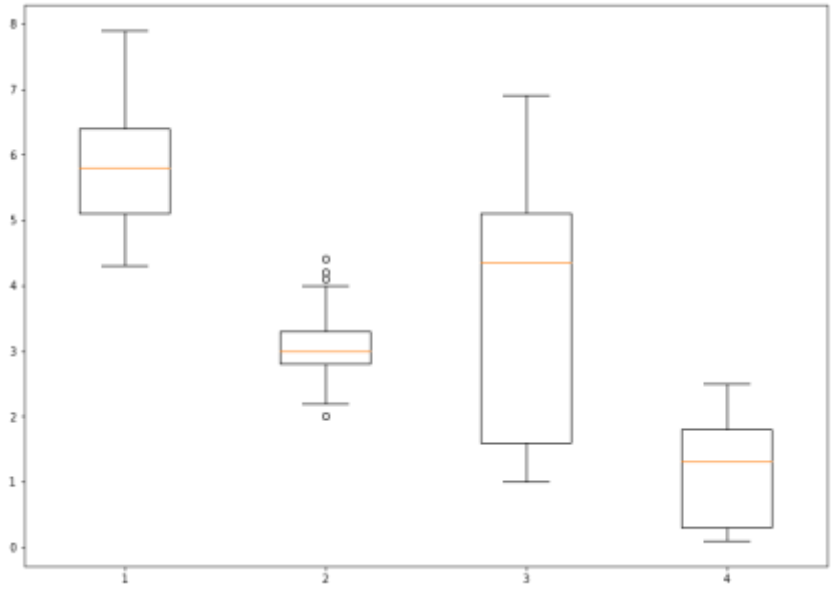
3. Hasil dan Pembahasan

Hasil percobaan yang dilakukan dapat dilihat pada Tabel 1. Pada tabel tersebut terdapat hasil akurasi dari 12 kali percobaan yang dihasilkan dari berbagai kombinasi dataset dan metode normalisasi data.

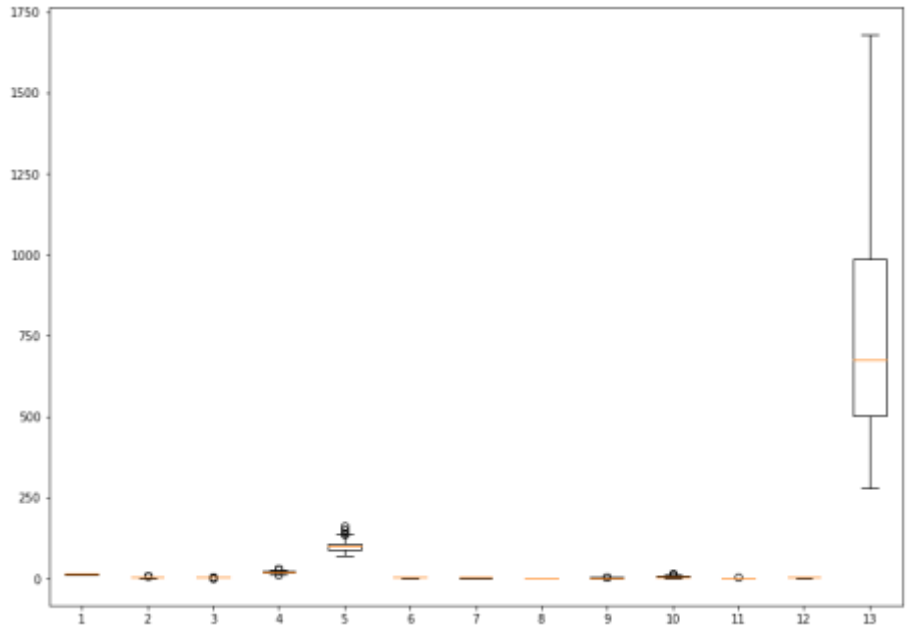
Tabel 1. Hasil Percobaan

No.	Dataset	Akurasi (%)			
		Tanpa Normalisasi	MinMax	Z-Score	MaxAbs
1	Iris	98.00	96.00	96.00	97.33
2	Wine	30.32	96.10	97.76	94.41
3	Breast cancer	62.92	97.19	97.90	94.91

Seperti yang terlihat pada Tabel 1, Dataset Iris tanpa normalisasi justru menghasilkan performa BP tertinggi dengan akurasi 80.00%. Performa terendah didapat ketika data di normalisasi dengan menggunakan MinMax Normalization dan Z-Score Normalization, dengan akurasi 96.00%. Penyebab Dataset Iris mendapatkan performa tertinggi ketika data tidak dinormalisasi bisa disebabkan karena rentang nilai antar atribut di dataset tersebut tidak terlalu jauh. Hal ini dapat terlihat pada *box-plot* Dataset Iris (Gambar 1).

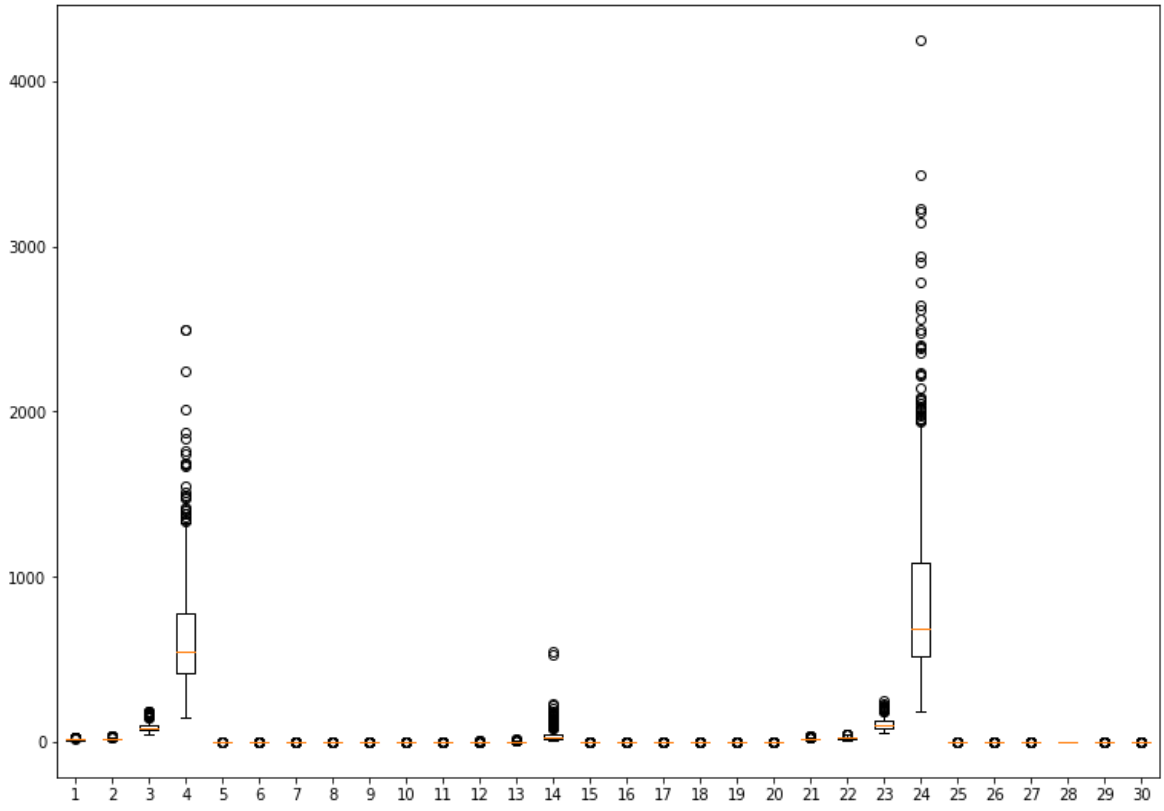


Gambar 1. *Box-Plot* Dataset Iris



Gambar 2. *Box-Plot* Dataset *Wine*

Pada Dataset *Wine* terlihat bahwa performa BP pada data yang dinormalisasi lebih baik dibandingkan dengan data yang tidak dinormalisasi (lihat Tabel 1). Performa tertinggi didapat pada data yang dinormalisasi menggunakan Z-Score *Normalization*, yaitu mendapatkan akurasi sebesar 97.76%. Sedangkan data tanpa normalisasi mendapatkan akurasi yang sangat rendah, yaitu hanya 30.32%. Hal ini bisa disebabkan oleh ada atribut pada dataset yang memiliki rentang nilai terlalu jauh. Hal tersebut dapat dilihat pada *box-plot* di Gambar 2. Pada Gambar 2 tersebut terlihat bahwa atribut ke 13 memiliki rentang nilai yang berbeda jauh dengan atribut lainnya.



Gambar 2. *Box-Plot* Dataset *Breast Cancer*

Sama dengan Dataset *Wine*, pada Dataset *Breast Cancer*, data yang dinormalisasi mendapatkan performa yang lebih baik dari data yang tidak dinormalisasikan (lihat Tabel 1). Akurasi tertinggi didapat dari data yang dinormalisasi dengan menggunakan Metode Z-*Score Normalization*, yaitu 97.90%. Sedangkan data tanpa normalisasi mendapatkan akurasi yang rendah, yaitu hanya 62.92%. Hal ini bisa disebabkan adanya atribut pada dataset ini yang memiliki rentang nilai jauh berbeda dari atribut lainnya. Seperti yang terlihat pada *box-plot* di Gambar 3, atribut ke 4 dan ke 24 memiliki rentang nilai yang berbeda jauh dari atribut yang lainnya.

4. Kesimpulan

Berdasarkan hasil percobaan pada tiga dataset dan tiga metode normalisasi dapat disimpulkan bahwa jika data yang digunakan antar atributnya tidak memiliki perbedaan rentang nilai yang terlalu jauh, maka penerapan Algoritma BP untuk klasifikasi tanpa normalisasi data bisa menjadi pilihan terbaik. Sedangkan jika dataset yang digunakan mempunyai atribut-atribut yang memiliki perbedaan rentang nilai yang besar, maka menggunakan normalisasi adalah pilihan terbaik. Hasil percobaan juga menunjukkan bahwa metode normalisasi terbaik adalah Z-*Score Normalization*.

Referensi

[1] Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised leaning. International journal of computer science, 1(2), 111-117.

[2] Mustaffa, Z., & Yusof, Y. (2011). A comparison of normalization techniques in predicting dengue outbreak. In International Conference on Business and Economics Research (Vol. 1, pp. 345-349).

[3] Jain, S., Shukla, S., & Wadhvani, R. (2018). Dynamic selection of normalization techniques using data complexity measures. Expert Systems with Applications, 106, 252-262.

[4] Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462.

[5] Vora, K., Yagnik, S., & Scholar, M. (2014). A survey on backpropagation algorithms for feedforward neural networks.

[6] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation. California Univ San Diego La Jolla Inst for Cognitive Science.

[7] Eesa, A. S., & Arabo, W. K. (2017). A normalization methods for backpropagation: a comparative study. Science Journal of University of Zakho, 5(4), 319-323.

[8] Hoang, A. T., Nižetić, S., Ong, H. C., Tarelko, W., Le, T. H., Chau, M. Q., & Nguyen, X. P. (2021). A review on application of artificial neural network (ANN) for performance and emission characteristics of diesel engine fueled with biodiesel-based fuels. Sustainable Energy Technologies and Assessments, 47, 101416.

[9] Pandey, A., & Jain, A. (2017). Comparative analysis of KNN algorithm using various normalization techniques. International Journal of Computer Network and Information Security, 9(11), 36.

[10] Raju, V. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020, August). Study the influence of normalization/transformation process on the accuracy of supervised classification. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 729-735). IEEE.

[11] Jayalakshmi, T., & Santhakumaran, A. (2011). Statistical normalization and back propagation for classification. International Journal of Computer Theory and Engineering, 3(1), 1793-8201.

[12] Pan, J., Zhuang, Y., & Fong, S. (2016, September). The impact of data normalization on stock market prediction: using SVM and technical indicators. In International Conference on Soft Computing in Data Science (pp. 72-88). Springer, Singapore.

[13] Folorunso, T. A., Aibinu, A. M., Kolo, J. G., Sadiku, S. O., & Orire, A. M. (2018). Effects of data normalization on water quality Model in a recirculatory aquaculture system Using artificial neural network.

[14] Chamidah, N., & Salamah, U. (2012). Pengaruh normalisasi data pada jaringan syaraf tiruan backpropagasi gradient descent adaptive gain (bpgdag) untuk klasifikasi. ITSMART: Jurnal Teknologi dan Informasi, 1(1), 28-33.

[15] Nayak, S. C., Misra, B. B., & Behera, H. S. (2014). Impact of data normalization on stock index forecasting. International Journal of Computer Information Systems and Industrial Management Applications, 6(2014), 257-269.

[16] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of eugenics, 7(2), 179-188.

[17] Forina, M. An extendible package for data exploration, classification and correlation (1998). Institute of Pharmaceutical and Food Analysis and Technologies.

[18] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer letters, 77(2-3),

163-171.
[19] Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. Operations Research, 43(4), 570-577.
[20] Izonin, I., Tkachenko, R., Shakhovska, N., Ilchyshyn, B., & Singh, K. K. (2022). A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain. Mathematics, 10(11), 1942.