An Investigation on Disparity Responds of Machine Learning Algorithms to Data Normalization Method

Haval A. Ahmed[1], Peshawa J. Muhammad Ali[1], Abdulbasit K. Faeq[1], and Saman M. Abdullah[1,2]

[1]Department of Software Engineering, Faculty of Engineering, Koya University, Koya KOY45, Kurdistan Region, F.R. Iraq

[2]Department of Computer Engineering, Faculty of Engineering, Tishk International University, Erbil, Kurdistan Region, F.R. Iraq

***Abstract*—Data normalization can be useful in eliminating the effect of inconsistent ranges in some machine learning (ML) techniques and in speeding up the optimization process in others. Many studies apply different methods of data normalization with an aim to reduce or eliminate the impact of data variance on the accuracy rate of ML-based models. However, the significance of this impact aligning with the mathematical concept of the ML algorithms still needs more investigation and tests. To identify that, this work proposes an investigation methodology involving three different ML algorithms, which are support vector machine (SVM), artificial neural network (ANN), and Euclidean-based K-nearest neighbor (E-KNN). Throughout this work, five different datasets have been utilized, and each has been taken from different application fields with different statistical properties. Although there are many data normalization methods available, this work focuses on the min-max method, because it actively eliminates the effect of inconsistent ranges of the datasets. Moreover, other factors that are challenging the process of min-max normalization, such as including or excluding outliers or the least significant feature, have also been considered in this work. The finding of this work shows that each ML technique responds differently to the min-max normalization. The performance of SVM models has been improved, while no significant improvement happened to the performance of ANN models. It is been concluded that the performance of E-KNN models may improve or degrade with the min-max normalization, and it depends on the statistical properties of the dataset.**

***Index Terms*—Min-max normalization, Support vector machine, Artificial neural network, Euclidean-based K-nearest neighbor, Mean squared error.**

I. Introduction

Min-max data normalization is one of the data scaling methods that cast data in a specific range of [0,1] or [−1,1]. The main aim of such scaling is improving the performance of machine learning (ML) techniques. Min-max normalization is mainly used to speed up the convergence of some techniques utilizing gradient descent algorithm for convergence, like in artificial neural networks (ANN), and to eliminate the domination of some features over others in the techniques using distance measures like Euclidean K-nearest neighbor (E-KNN). Supposing that a dataset may contain different feature ranges, distance measures like Euclidean may assign more weight to features with larger ranges than those with small ranges. Therefore, min-max data normalization is used to equalize the weight of these features and make them have the same effect on the decision-making process. Nevertheless, there are still arguments among researchers about the exact impact of the min-max data normalization on the performance of supervised ML techniques, some of them look at normalization as a necessary step in machine learning process, while other see it as unnecessary step. The causes of this difference may include the quality of the datasets, the nature of the dataset, the application field, or to the machine learning technique itself and how it deals with the data. On the other hand, there are many different techniques of data normalization, they may respond differently to the different ML techniques. Min-max is one of the most used techniques with different ranged attribute datasets. It's easy to implement and has an approved effects on the performance of the models. This study aims to reveal the ambiguity of the real effect of the min-max data normalization and more specifically to investigate its impacts on the regression performance of ML models. Therefore, the main questions that addressed by this work is why ML techniques show disparity responds to the min-max normalization method.

The rest of this article is structured as follows: Section 2 presents the related works, whereas Section 3 is the methodology used in this research work consisting of six subsequent stages: Selecting and downloading the datasets, implementing min-max feature normalization, implementing ML techniques on the datasets, removing outliers, feature selection, and removing outliers with feature selection. Section 4 summarizes all the observed results, and Section 5 discusses the observed results. Section 6 concludes this research work.

ARO-The Scientific Journal of Koya University
Vol. X, No. 2 (2022), Article ID: ARO.10970. 9 pages
DOI: 10.14500/aro.10970
Received: 26 April 2022; Accepted: 28 August 2022
Regular research paper: Published: 19 September 2022
Corresponding author's e-mail: saman.mirza@koyauniversity.org

II. Related Works

The majority of the works in the literature that have investigated the effect of the min-max data normalization reported a positive impact of the min-max normalization on the adopted ML techniques in their studies (Dadzie and Kwakye, 2021; Shahriyari, 2017), while some other studies determined that its usefulness varies from good to bad depending on the nature of the datasets and the ML model (Ambarwari, Adrian, and Herdiyeni, 2020) (Ahsan, et al., 2021). On the other hand, very limited studies concluded the degradation of the ML model accuracy with the present of min-max normalization. However, no comprehensive interpretation about their achieved results has been mentioned in those studies (Singh, et al., 2015).

In this work, we used most datasets from the most recognized and benchmarked dataset repository available online which is the University of California Irvine (UCI) Repository Dataset. We found that most of the researcher works in this area using datasets from the same repository (Ahsan, et al., 2021; Bhardwaj, Mishra, and Desikan, 2018; Dadzie and Kwakye, 2021; Jayalakshmi and Santhakumaran, 2011; Pires, et al., 2020; Shorman, et al., 2018). The research direction is progressed toward determining the effects of the min-max scaling regardless of the nature of the dataset application. Therefore, in this work, five different benchmarked datasets from the UCI repository have been adopted as well having different number of records and features. Because this work focuses on three ML techniques, which are SVM, ANN, and E-KNN, reviewing some relevant works on ML techniques, in general, and on these three techniques, more specifically, are needed.

One of the most common ML techniques is SVM. Many research works showed that SVM has more sensitive responds than ANN and E-KNN toward the normalization techniques. Results in many SVM-based works showed the usefulness of using min-max normalization with SVM (Dadzie and Kwakye, 2021; Shahriyari, 2017; Ambarwari, Adrian, and Herdiyeni, 2020). Despite that, there are still some studies proved that normalization has no effect or has very little effect on the accuracy rate of the SVM-based models (Singh, et al., 2015). Moreover, there are studies (Ahsan, et al., 2021) proved the degradation of the performance of the SVM-based models while attached to min-max normalization method.

On the other hand, research works commented differently on the suitability of using min-max method with ANN. Some works showed the negative effect of min-max normalization (Singh, et al., 2015), whereas others concluded that no significant performance improvements were observed (Jayalakshmi and Santhakumaran, 2011). No improvements were also observed for higher-order neural networks (Prasetyo, Setiawan, and Adji, 2020) and deep learning algorithms (Pires et al., 2020), while other research works reported an obvious improvement in the accuracy of the models (Ambarwari, Adrian, and Herdiyeni 2020).

The ambiguity is not only existed with ANN-based works, the uncertainty about the suitability of the min-max method and its impact on the accuracy of ML-based models also found in E-KNN-based works that used for classification or clustering. Some studies observed that attaching min-max with E-KNN technique will improve the accuracy rate within a very small range (Ambarwari, Adrian, and Herdiyeni 2020; Dadzie and Kwakye, 2021), however, other studies showed a significant improvement in the accuracy rate of the E-KNN-based models (Ahsan, et al., 2021; Rajeswari and Thangavel, 2020). There other research works stating the min-max normalization impact on the performance of E-KNN depends on the nature of the dataset, it may enhance or degrade the performance of E-KNN models (Muhammad Ali, 2022).

It becomes clear from reviewing the above relevant works that there is non-clear vision about the situation where the min-max normalization could be utilized with ML techniques for accuracy improvement. In other words, the question about the condition(s) or the circumstance(s) that make the min-max normalization responds a positive or negative effect on the performance of supervised learning models, needs be answered. To the best of our knowledge, we could not find a comprehensive study that tackles this problem.

Whereas the focus of this study is on evaluating the impact of the min-max normalization on the performance of some adopted regression models; it is necessary to review the validation measurements that have been used in the previous works. For regression models, the mean squared error (MSE) is the widely adopted performance measurement (Rajeswari and Thangavel, 2020; Singh, et al., 2015; Jayalakshmi and Santhakumaran, 2011; Shorman, et al., 2018; Bhardwaj, Mishra, and Desikan 2018). Moreover, the root mean square error (RMSE) (Prasetyo, Setiawan, and Adji, 2020), the coefficient of determination (R2) (Aksu, Güzeller, and Eser, 2019), and mean absolute error (MAE) were also used as performance measurements. On the other hand, the fitting time is an important performance measurement which is the time needed to fit the models (Shahriyari, 2017). The number of steps for convergence or number of iterations, epochs, and the complexity level of the model are other measures that could be used as performance measure. In this research work, performance measurement depends on the MSE.

Beside the mathematical process of ML techniques, data characteristics are another problem that challenging the performance of min-max method for accuracy impairment. These challenges are “outliers,” “noise amplification,” and “out of range data.” The min-max normalization preserves the real relations among instances of the same feature, which makes it very sensitive to “outliers.” Having anomalies in any feature forces the data to aggregate in a small range between 0 and 1, and leaves a wide range empty, this makes anomalies study a necessary step before implementing min-max normalization (Kappal, 2019). Moreover, the min-max eliminates the real differences among different features, this is called noise amplification, which is the enlargement of the small effect attributes and making them equal to the big effect attributes, which leads to a decrease in the accuracy of the models (Pires, et al., 2020). This challenge could be solved by removing the features with little significance among the input features of the models.

In this work, an outlier study and feature selection study were implemented on the original dataset. There are many methodologies available for outlier removing, we utilized interquartile range method. Literatures utilized different feature selection methods (Sattari, et al., 2021); this work used Pearson correlation method. Therefore, four different pairwise result sets are adopted: The pairwise result achieved using the original dataset, the pairwise result of the outlier clean dataset, the pairwise result after removing the least significant feature, and finally, the pairwise result set after implementing both outlier and features selection on the original dataset.

The scaling process of the min-max normalization depends on the features' minimum and maximum values, therefore, an "out of range" problem arises when the trained model is receiving a data instance outside the feature bounds, which leads to wrong predictions. The "out of range" challenge mostly happens with time series datasets. Studies suggested many solutions in this regard (Ogasawara, et al., 2010), widening the range between minimum and maximum values by a specific ratio of 20% is one of the suggested solutions. This free space above the max and below the min works as spare space in case such data appeared, but the time series data are beyond the scope of this article.

From the above literature review, we can conclude that there is a clear disagreement among researchers about the impacts of min-max data normalization on the regression performance of the ML models. Utilizing data normalization always enhances the performance of the ML model. In our opinion, this question needs more investigating before answering it. This research aims to reveal the ambiguity of the impact of the min-max data normalization on the regression performance of three ML algorithms SVM, ANN, and E-KNN by testing five different datasets and comparing the results with and without data normalization.

III. Methodology

In this research, three different ML techniques have been tested on five benchmark datasets to investigate the impact of min-max data normalization on the regression performance of these ML techniques. The methodology can be summarized as follows, Fig. 1.

Step 1: Adopting suitable datasets from the UCI website,
Step 2: Implementing min-max feature normalization, range [0,1],
Step 3: Applying and validating ML techniques on both original and the normalized datasets,
Step 4: Comparing the MSE results of the both cases to show the impact of min-max on the performance of the models,
Step 5: Implementing anomaly detection and removing outliers to the original datasets and then repeating Steps 2, 3, and 4,
Step 6: Implementing feature selection to the original datasets and then repeating Steps 2, 3, and 4
Step 7: Implementing anomaly detection and feature selection together to the original datasets and then repeating Steps 2, 3, and 4.

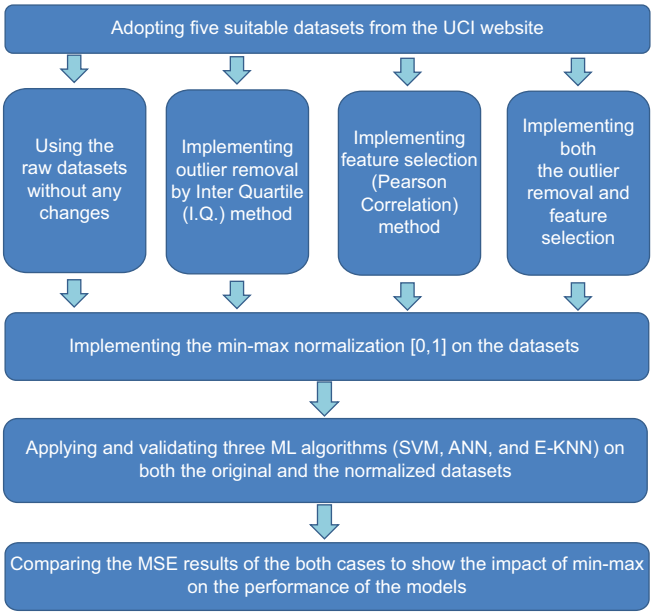

Fig. 1. The methodology of the research work.

A. The UCI datasets

Five different benchmark datasets were downloaded from the machine learning repository website of the University of California Irvine (UCI) (Dheeru and Graff, 2019). The reason behind selecting these datasets is the existing variation in the ranges of records among all attributes. Some of the datasets have big differences in their ranges like in the Airfoil Self-Noise Dataset (Table I) or very similar ranges like in power plant dataset (Table II). Such a variation is expected to have an impact on the regression performance of ML algorithms, and consequently, the impact of the min-max scaling method will appear more, which is the main target that this work aimed to investigate. The datasets belong to the real applications of physics, life sciences, engineering, and business, Table III. The dataset's statistical properties are shown in Tables I,II,IV-VI.

B. Implementing min-max feature normalization, range [0,1].

As shown in Equation (1), a normalized data sample x' could be obtained from the original data sample x. For an attribute, it is mostly dependent on instances with the maximum and minimum values in the same attribute. In this normalization method, the original data sample component values will be transformed to the range of [0,1].

$$x' = [\ (\frac{x - \text{oldMin}}{\text{oldMax} - \text{oldMin}}) * (\text{newMax} - \text{newMin})\ ] + \text{newMin} \quad (1)$$

Where:
x' is the normalized data sample,
x is the original data sample,
oldMin is the minimum data among any attribute of the original dataset,
oldMax is the maximum data among any attribute of the original dataset,
newMin is the minimum of the normalized dataset, and
newMax is the maximum of the normalized dataset.

Table I
Statistical Properties of the Airfoil Self-Noise Dataset (Brooks, Pope, and Marcolini, 1989)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	20,000	200	19800	2886.380572	9,938,717.384	3152.573137
F2	22.2	0	22.2	6.782302063	35.0242405	5.918128125
F3	0.3048	0.0254	0.2794	0.136548237	0.008749868	0.093540728
F4	71.3	31.7	39.6	50.86074518	242.5116138	15.5727844
F5	0.0584113	0.000400682	0.058010618	0.01113988	0.000172929	0.013150234
Target	140.987	103.38	37.607	124.8359428	47.59146318	6.898656622

SD: Standard deviation

Table II
Statistical Properties of the Combined Cycle Power Plant Dataset (Tüfekci, 2014)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	37.11	1.81	35.3	19.65123119	55.53935724	7.45247323
F2	81.56	25.36	56.2	54.30580372	161.4905445	12.707893
F3	1033.3	992.89	40.41	1013.259078	35.2691519	5.938783706
F4	100.16	25.56	74.6	73.30897784	213.1678478	14.60026876
Target	495.76	420.26	75.5	454.3650094	291.2823183	17.066995

SD: Standard deviation

Table III
Datasets Properties

Dataset	Number of features	Type of features	Number of instances	Supervised learning	Application
Airfoil Self-Noise Dataset (Brooks, Pope, and Marcolini, 1989)	5	Real	1503	Regression	Physics
Physicochemical Properties of Protein Tertiary Structure Dataset (Rana, 2013)	9	Real	9146 instances included	Regression	Life sciences
Combined Cycle Power Plant Dataset (Tüfekci, 2014)	4	Real	9568	Regression	Energy
Concrete Compressive Strength Dataset (Yeh, 1998)	8	Real	1030	Regression	Civil engineering
Real Estate Valuation Dataset (Yeh and Hsu, 2018)	6	Integer and Real	414	Regression	Business

Table IV
Statistical Properties of the Physicochemical Properties of Protein Tertiary Structure Dataset (Rana, 2013)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	32,240.2	2783.15	29,457.05	9873.68162	16,094,604.52	4011.808135
F2	11,787.1	403.5	11,383.6	3016.435929	2,102,621.452	1450.041879
F3	0.56848	0.09362	0.47486	0.302155567	0.003941913	0.062784658
F4	343.239	10.6891	332.5499	103.4039974	3018.359088	54.9395949
F5	4,467,324.7	374,315.5155	4,093,009.223	1,369,092.965	311,794,123,479.2	558,385.2823
F6	470.897	33.6462	437.2508	145.5447009	4803.146285	69.30473494
F7	83,153.57	1108.9	82,044.67	3987.14593	3,536,332.356	1880.513854
F8	337	0	337	70.04286027	3192.870115	56.50548747
F9	47.4559	15.5049	31.951	34.48790348	35.17094729	5.930509868
Target	20.981	0	20.981	7.833154384	37.46611427	6.120956974

SD: Standard deviation

Table V
Statistical Properties of the Concrete Compressive Strength Dataset (Yeh, 1998)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	540	102	438	281.1656311	10,921.74265	104.5071416
F2	359.4	0	359.4	73.89548544	7444.083725	86.27910364
F3	200.1	0	200.1	54.18713592	4095.548093	63.99646938
F4	247	121.75	125.25	181.5663592	456.0602447	21.35556707
F5	32.2	0	32.2	6.20311165	35.6826025	5.973491651
F6	1145	801	344	972.9185922	6045.656228	77.75381809
F7	992.6	594	398.6	773.5788835	6428.099159	80.1754274
F8	365	1	364	45.66213592	3990.437729	63.16991158
Target	82.599225	2.331807832	80.26741697	35.81783583	279.0797167	16.70567917

SD: Standard deviation

Table VI
Statistical Properties of the Real Estate Valuation Dataset (Yeh and Hsu, 2018)

Feature	Maximum value	Minimum value	Range	Mean	Variance	SD
F1	2013.5833	2012.666667	0.9166666	2013.148953	0.079521365	0.281995327
F2	43.8	0	43.8	17.71256039	129.7887038	11.39248453
F3	6488.021	23.38284	6464.63816	1083.885689	1,592,920.631	1262.109595
F4	10	0	10	4.094202899	8.676334351	2.945561806
F5	25.01459	24.93207	0.08252	24.96903007	0.000154013	0.012410197
F6	121.56627	121.47353	0.09274	121.5333611	0.000235536	0.015347183
Target	117.5	7.6	109.9	37.98019324	185.1365075	13.6064877

SD: Standard deviation

C. Applying and validating ML techniques on the datasets

In this work, SVM, ANN, and E-KNN were adopted as regression methods. The target of applying all tests was to determine the impact of min-max data normalization on the regression performance of these techniques through implementing the technique on the original dataset and the normalized version individually then comparing their results. The reason for adopting SVM, ANN, and E-KNN techniques is the diversity of their nature. SVM utilizes Lagrange optimizer and tries to maximize the margins using different kernel functions whereas ANN utilizes a gradient descent algorithm as an optimization technique to minimize errors and to reach the goals faster. On the other hand, E-KNN utilizes Euclidean distance to determine the distance between the tested samples with the neighbors, where this distance is affected by the range of the values of the features. In all the tests, mean squared error (MSE) is computed as a performance measure and 10-fold as a validation technique.

For testing the SVM on all of the different datasets, four different kernel functions were considered (linear, Gaussian, radial base, and polynomial) on both the original and the normalized datasets. The MSE of the test set has been used for comparison purposes. The tests were designed such that the code runs on both the original datasets and the normalized dataset with the same parameters to determine the impact of the min-max normalization in MSE. The minimum MSE of the four tests was considered for comparison purposes with other techniques.

For each one of the datasets, 50 different models of ANN were tested. These models include either one or two hidden layers, with a different number of nodes in each hidden layer ranging between 3 and 100 nodes. In addition, different transfer functions were used in the hidden layers (sigmoid, tanh, and ReLU). The same model with the same parameters has been applied on both datasets (the original dataset and the normalized dataset), where the minimum MSE of the 50 tests was considered for comparison purposes with other techniques. It is known that ANN model results vary from one run to another, therefore, each one of the 50 different models is tested 10 times and the average is presented as the MSE result of the model within an acceptable standard deviation range. This is despite that each test of the 10 tests was validated by 10-fold validation.

A group of 33 models that used E-KNN (23 weighted-neighbor models and 10 traditional E-KNN models) was tested on the five datasets. The weighted group consisted of different models having five neighbors with different contribution weights ranging from 5% to 100%. The best weighted E-KNN model for all the datasets was 50%, 20%, 15%, 10%, and 5% from the closest neighbor to the farthest neighbor accordingly. The second group consisted of different E-KNN models considering up to 10 neighbors with equal contribution weights. The tests were implemented on both datasets, the original and the normalized. Minimum MSE among all the adopted experiments is observed to be used for comparison purposes.

D. Implementing anomaly detection and removing outliers

An outlier detection is conducted on the adopted five datasets by implementing two different methods, mean-standard deviation, and the interquartile method. The first method is considering the data lay outside the range (mean ± 3 × standard deviation) as an anomaly, and the second method is considering the data lay outside the range (Q1 - 3 × IQ, Q3 + 3 × IQ) as an anomaly, (Table VII).

In this research work, the wider range is considered for anomaly detection to decrease the number of anomalies, which is the interquartile method. Any data laid outside the wider range had been removed from the datasets. The same previous ML techniques were simply repeated, and the results were observed for comparison purposes.

E. Implementing feature selection

To determine the least significant feature among the features of each of the five datasets, this research adopts Pearson correlation. The correlation values are ranging between −1 and 1, the least significant feature has the correlation value between the feature and the target closer to zero (Table VIII).

F. Implementing anomaly detection and feature selection together

In this step, the combined effect of feature selection and outlier removal is investigated by implementing both procedures in the previous steps and then repeating the same above procedure. This step shows the impact of anomaly removal and the least significant feature removal together.

IV. Results

The minimum MSE results using different machine learning techniques on the five datasets are shown in Table IX and Fig. 2.

 http://dx.doi.org/10.14500/aro.10970

After removing the outliers from the datasets, the MSE values changed in some of the datasets as one can observe in Table X and Fig. 3.

The same techniques with the same parameters were repeated after removing the least significant feature among all other features of each dataset. The results are shown in Table XI and Fig. 4.

Again, and after removing both the outliers and the least significant feature, the same techniques with the same parameters were repeated in Table XII and Fig. 5.

Table VII
Number of Outliers in the Adopted Datasets using (Mean ± 3x Standard Deviation) and (Q1 − 3 × IQ, Q3 + 3 × IQ) Methods

Dataset	Number of outliers (mean ± 3 × SD)	Number of outliers (Q1 − 3 × IQ, Q3 + 3 × IQ)
Airfoil self-noise dataset	78	35
Physicochemical properties of protein tertiary structure dataset	806	213
Combined cycle power plant dataset	58	0
Concrete compressive strength dataset	49	33
Real estate valuation dataset	10	8

SD: Standard deviation

Table VIII
The Pearson Correlation Values for the Least Significant Features of the Adopted Datasets with the Target Feature

Dataset	Least significant feature	Pearson correlation between the feature and the target
Airfoil self-noise dataset	F4	0.1251
Physicochemical properties of protein tertiary structure dataset	F7	−0.0018
Combined cycle power plant dataset	F4	0.3898
Concrete compressive strength dataset	f3	−0.1058
Real estate valuation dataset	F1	0.0875

V. Discussion

The MSE values in the four (Tables IX-XII) are the minimum MSE observed after implementing a large number of various models and structures validated by 10-fold validation. For SVM, each number is the minimum of four models with different kernels (linear, Gaussian, radial base, and polynomial), which are applied to the original and the normalized data separately resulting in 8 models per case and 32 models for all of the cases for each dataset, 160 models for all the five datasets. Whereas for ANN, the parameters are the number of hidden layers (up to two layers), the number of nodes in each layer (up to 100 nodes), and the type of activation function (sigmoid, tanh-hyperbola, and ReLU). Totally, 50 models were applied to the original and the normalized data separately, resulting in 100 models for each case, and a total of 400 models per each one of the five datasets, 2000 ANN models were tested as a grand total. Finally, regarding the E-KNN pair 66 models were tested, the total is 264 for each pair for each dataset, the grand total is 1320 models which were tested.

For the SVM, it is clear that the adopted min-max normalization shows a positive impact on the results observed from all the datasets. There are differences in the ratio of the improvement, but we can see clearly that all the datasets responded positively to the min-max normalization with SVM (Table IX). This improvement in the results is reported in the literature as well (Ambarwari, Adrian, and Herdiyeni 2020; Dadzie and Kwakye, 2021; Shahriyari, 2017). The reason is due to the fact that the SVM does not have any tool to weigh one dimension versus other(s), but rather it focuses on optimizing the line, plane, or the hyperplane that separates the classes.

The results for ANN are not similar to the SVM, as there is no significant difference observed in comparing the pairwise results of the five datasets. The reason behind this is the existence of transfer functions (activation functions) in each hidden layer node of the neural network, where they may

Table IX
Minimum Values of the Mean Squared Error Results of the Different Machine Learning Models (Original *vs*. Minimum-Maximum Normalized Datasets)

Serial number	Machine learning technique	Airfoil self-noise dataset		Physicochemical dataset		Power plant dataset		Concrete strength dataset		Real estate valuation dataset	
		Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization
1	SVM	48	14	37	25	21	17	118	45	101	60
2	ANN	3	3	21	21	15	15	28	29	63	63
3	E-KNN	35	4	38	18	15	13	60	65	64	70

SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor

Table X
Minimum Values of the Mean Squared Error Results of the Different Machine Learning Models (After Removing Outliers)

Serial number	Machine learning technique	Airfoil self-noise dataset		Physicochemical dataset		Power plant dataset		Concrete strength dataset		Real estate valuation dataset	
		Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization
1	SVM	45	13	37	25	21	17	89	34	97	60
2	ANN	3	2	20	20	15	15	31	31	63	66
3	E-KNN	34	4	38	18	15	13	61	52	66	70

SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor

Table XI
Minimum Values of the Mean Squared Error Results of the Different Machine Learning Models (After Removing the Least Significant Feature)

Serial number	Machine learning technique	Airfoil self-noise dataset		Physicochemical dataset		Power plant dataset		Concrete strength dataset		Real estate valuation dataset	
		Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization
1	SVM	28	16	37	25	23	19	125	48	99	63
2	ANN	4	4	21	21	16	16	37	33	66	62
3	E-KNN	13	4	38	18	14	15	63	66	65	68

SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor

Table XII
Minimum Values of the Mean Squared Error Results of the Different Machine Learning Models (After Removing Outliers and Least Significant Feature)

Serial number	Machine learning technique	Airfoil self-noise dataset		Physicochemical dataset		Power plant dataset		Concrete strength dataset		Real estate valuation dataset	
		Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization	Original	Normalization
1	SVM	25	15	37	25	22	19	95	40	96	63
2	ANN	4	4	21	20	16	16	35	35	63	64
3	E-KNN	12	5	38	19	14	15	64	52	64	66

SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor

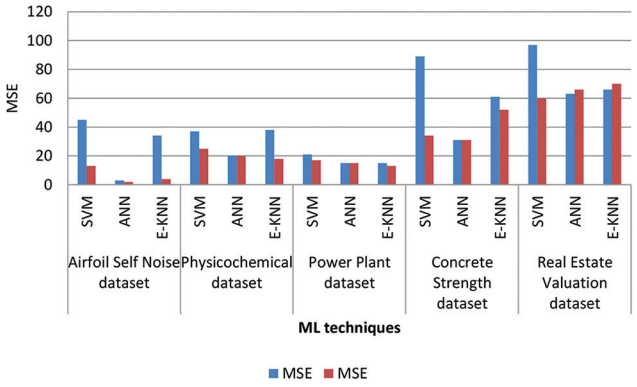

Fig. 2. Minimum values of the mean squared error results of the different machine learning models (original vs. min-max normalized datasets).

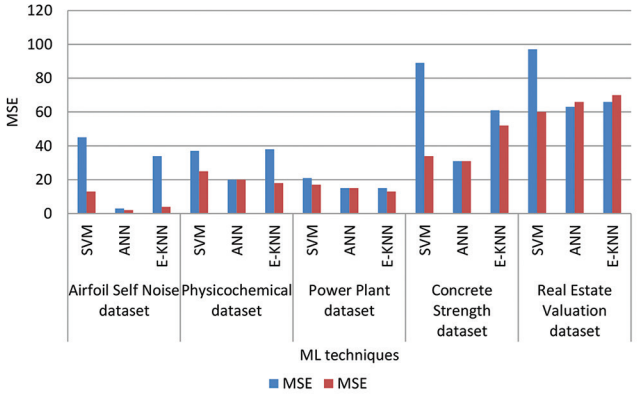

Fig. 3. Minimum values of the mean squared error results of the different machine learning models (after outlier removal).

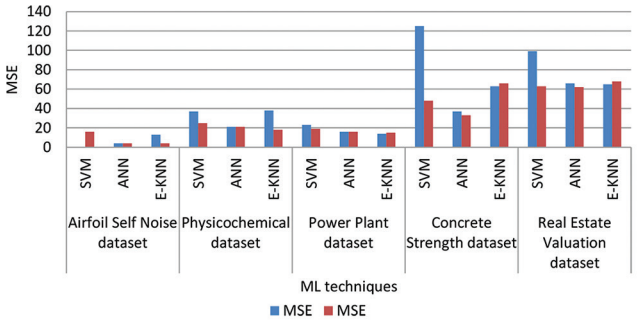

Fig. 4. Minimum values of the mean squared error results of the different machine learning models (after removing the least significant feature).

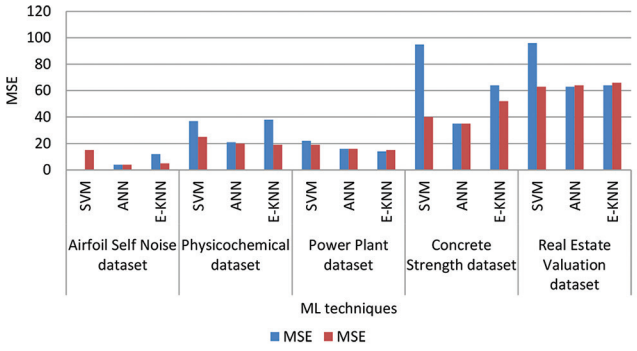

Fig. 5. Minimum values of the mean squared error results of the different machine learning models (after removing outliers and least significant feature).

work like a normalization layer as well. The backpropagation procedure to adjust weights performs a denormalization process of the data; this is happening in each training cycle during the training session of the ANN models (Table IX).

For E-KNN, the results are different because the impact of min-max depends on the nature of the dataset itself. The MSE results of two of the datasets (Airfoil Self-Noise and Physicochemical) have been improved significantly, whereas no improvement happened to the power plant dataset after implementing the min-max data normalization, the reason could be because of the close ranges of the various dimensions in this dataset, while the MSE results of the two datasets concrete-strength and house-valuation datasets are degraded significantly. The reason behind the diversity of the effect of min-max data normalization on the MSE results of E-KNN returns to its dependence on the Euclidean distance

Table XIII
Comparing the Reliability of this Research with Other Previous Literatures

Serial number	Research work	Performance of SVM model	Performance of ANN model	Performance of E-KNN model	Number of datasets used	Type of application of the datasets
1	This research work	Improvement	No significant improvement observed	Depends on the nature of the dataset	5	5
2	Dadzie and Kwakye 2021	Improvement	N/A	No significant change	1	1
3	Shahriyari 2017	Improvement	Degradation	N/A	1	1
4	Ambarwari, et al., 2020	Improvement	Improvement	No significant change	1	1
5	Ahsan, et al., 2021	Degradation	N/A	Improvement	1	1
6	Singh, et al., 2015	No significant change	Degradation	N/A	1	1
7	Jayalakshmi and Santhakumaran, 2011	N/A	No significant change	N/A	1	1
8	Rajeswari and Thangavel, 2020	N/A	N/A	Improvement	5	1

SVM: Support vector machine, ANN: Artificial neural network, E-KNN: Euclidean-based K-nearest neighbor, N/A: Not available

which will be affected significantly by the normalization process in a positive or a negative way (Table IX).

After implementing the anomaly detection process to the five datasets and removing all the outliers according to the procedures mentioned in the methodology section of this article (Table VII), generally, and similar to the previous experiments, SVM has been improved by implementing min-max normalization, no significant improvement was observed for ANN, whereas the E-KNN still depends on the nature of the datasets. The normalized data of the concrete-strength dataset showed a little bit more positive response by implementing SVM and E-KNN to the anomaly clean datasets, whereas the ANN model remained stable. This is showing the computational power of ANN against anomalies, (Table X).

Furthermore, feature selection is implemented by removing the least significant feature among all the features of the dataset. A drawback of removing the least significant feature is not always significantly increasing the performance of the regression model, especially in the power plant dataset (Table VIII). In general, the pairwise comparisons of the results in Table XI show the same conclusions observed with the original datasets (with all features), SVM improvement, no significant change to the ANN, and E-KNN depending on the nature of the dataset. The normalized result of implementing E-KNN on the real estate dataset is improved slightly (Table XI).

The combined effect of implementing both the removal of anomalies and removal of the least significant feature is shown in Table XII. The results of the normalized dataset of implementing E-KNN to the concrete-strength and real estate datasets are improved (Table XII).

The results in Tables IX-XII showed that the E-KNN is sensitive to the nature of the data, therefore, it is better to check the performance of the E-KNN models with and without normalization, with and without outliers, with and without feature selection, and then to decide which one is the best, as there is no specific rule could be generalized for E-KNN.

The power plant dataset (Table II) is not impacted by the min-max normalization during this research, and it did not contain outliers. By an intensive look at the feature ranges, we can see that it has very similar feature ranges, the standard deviation of the ranges is 15, which is a very small value compared to other datasets. This type of dataset is called a homogeneous dataset, it cannot be impacted by min-max data normalization.

Comparing the results observed in this research work with the results collected from reviewed literature showed that this research work implemented a better research methodology and analysis, also, the results are more precise and accurate (Table XIII). Most of the previous literature depended on testing only one dataset or in the best cases depended on using different datasets of one application field. This research work depended on five datasets belonging to five different real applications with different numbers of attributes and different attribute ranges. Therefore, the conclusions of this research are considered more reliable.

VI. Conclusion

The importance of implementing min-max data normalization mainly depends on the ML technique and the nature of the dataset. Experiments that are adopted in this work show that the min-max normalization is useful with SVM, whereas it makes no significant effect with ANN, which is because of the ANN's ability during the training stage to perform the normalization implicitly by itself. The powerful computation nature of ANN eliminates the effect of the min-max data normalization because it implicitly includes activation functions that work like the normalization layer. Even it eliminates the effect of outliers and least significant feature as well. Therefore, it is not important if min-max normalization is used with ANN or not. Depending on the nature of the dataset, min-max data normalization with E-KNN may result in performance improvement or degradation.

KNN uses distance measurements for determining the closest neighbors. In this research, Euclidean-based KNN is tested only. It has been concluded that using min-max data normalization may have a bad impact on the performance of the E-KNN, because it may eliminate the natural domination relations among the attributes and the target which leads to performance degradation, this has happened with two datasets concrete-strength and house-valuation. Therefore, it is better to test the E-KNN model with both the original dataset and

the normalized dataset before deciding if the min-max data normalization is useful or not.

E-KNN performance with min-max normalization may be improved by implementing outlier cleaning methods or by removing the least significant features, but this is not doing very well with SVM or ANN because of their powerful computation nature. No need to implement min-max data normalization to the homogeneous datasets whatever is the ML algorithm.

This research work has been compared to the literature in terms of the number of used datasets and their application fields. Unlike other works that adopt the use of a single dataset or datasets from one single application field, our research tried to be more reliable using five different datasets from five various application field.

References

Ahsan, M., Mahmud, M.A., Saha, P.K., Gupta, K.D. and Siddique, Z. 2021. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, 9, p.52.

Aksu, G., Güzeller, C.G. and Eser, M.T. 2019. The effect of the normalization method used in different sample sizes on the success of artificial neural network model. *International Journal of Assessment Tools in Education*, 6, pp.170-92.

Ali, P.J.M., 2022. Investigating the Impact of min-max data normalization on the regression performance of K-nearest neighbor with different similarity measurements. *ARO The Scientific Journal of Koya University*, 10, p.10955.

Ambarwari, A., Adrian, Q.J. and Herdiyeni, Y. 2020. Analysis of the effect of data scaling on the performance of the machine learning algorithm for plant identification. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 4, pp.117-122.

Bhardwaj, C.A., Mishra, M. and Desikan, K. 2018. Dynamic Feature Scaling for K-Nearest Neighbor Algorithm.

Brooks, T.F., Pope, D.T. and Marcolini, M.A. 1989. *Airfoil Self-noise and Prediction (NASA Reference Publication)*. In: Technical Report 1218. National Aeronautics and Space Administration, United States.

Dadzie, E. and Kwakye, K. 2021. Developing a Machine Learning Algorithm-Based Classification Models for the Detection of High-Energy Gamma Particles.

Dheeru, D. and Graff, C. 2019. UCI Machine Learning Repository. School of Information and Computer Science. Vol. 25. University of California, Irvine, CA, p27.

Jayalakshmi, T. and Santhakumaran, A. 2011. Statistical normalization and back propagationfor classification. *Journal of Computer Theory and Engineering*, 3 pp.89-93.

Kappal, S. 2019. Data normalization using median median absolute deviation MMAD based Z-Score for robust predictions vs. Min-max normalization. *London Journal of Research in Science Natural and Formal*, 19, pp.39-44.

Ogasawara, E., Martinez, L.V., De Oliveira, D., Zimbrão, G., Pappa, G.L. and Mattoso, M. 2010. Adaptive Normalization: A novel Data Normalization Approach for Non-stationary Time Series. In: *The 2010 International Joint Conference on Neural Networks (IJCNN)*, pp.1-8.

Pires, I.M., Hussain, F., Garcia, N.M., Lameski, P. and Zdravevski, E. 2020. Homogeneous data normalization and deep learning: A case study in human activity classification. *Future Internet*, 12, pp.194.

Prasetyo, J., Setiawan, N.A. and Adji, T.B. 2020. Improving normalization method of higher-order neural network in the forecasting of oil production. In: EDP Sciences.

Rajeswari, D. and Thangavel, K., 2020. The performance of data normalization techniques on heart disease datasets. *International Journal of Advanced Research in Engineering and Technology*, 11, pp.2350-2357.

Rana, P.S. 2013. Physicochemical properties of protein tertiary structure data set. UCI Machine Learning Repository, pp. Available from: https://www.archive.ics. uci. edu/ml/datasets/Physicochemical+Properties+ of+Protein+ Tertiary+ Structure. [Last accessed 2022 Apr 01].

Sattari, M.A., Roshani, G.H., Hanus, R., Nazemi, E., 2021. Applicability of time-domain feature extraction methods and artificial intelligence in two-phase flow meters based on gamma-ray absorption technique. *Measurement*, 168, p.108474.

Shahriyari, L. 2017. Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Brief Bioinformation*, 20, pp.985-94.

Shorman, A.R., Faris, H., Castillo, P.A., Merelo, J.J. and Al-Madi, N. 2018. The Influence of Input Data Standardization Methods on the Prediction Accuracy of Genetic Programming Generated Classifiers. *IJCCI 2018-Proceedings of the 10th International Joint Conference on Computational Intelligence,* pp.79-85.

Singh, B.K., Raipur, N.I.T., Verma, K. and Thoke, A.S. 2015. Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification. *International Journal of Computer Applications*, 116, pp.11-15.

Tüfekci, P. 2014. Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power and Energy Systems*, 60, pp.126-40.

Yeh, I.C. 1998. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, 28, pp.1797-1808.

Yeh, I.C. and Hsu, T.K. 2018. Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65, pp.260-271.