

Article

A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain

Ivan Izonin ^{1,*}, Roman Tkachenko ², Nataliya Shakhovska ¹, Bohdan Ilchyshyn ¹ and Krishna Kant Singh ³

¹ Department of Artificial Intelligence, Lviv Polytechnic National University, 79013 Lviv, Ukraine; nataliya.b.shakhovska@lpnu.ua (N.S.); bohdan.ilchyshyn.knm.2018@lpnu.ua (B.I.)

² Department of Publishing Information Technologies, Lviv Polytechnic National University, 79013 Lviv, Ukraine; roman.o.tkachenko@lpnu.ua

³ Department of Computer Science and Engineering, Jain (Deemed to Be University), Bangalore 560069, India; krishnaiitr2011@gmail.com

* Correspondence: ivan.v.izonin@lpnu.ua; Tel.: +380-98-888-96-87

Abstract: Data normalization is a data preprocessing task and one of the first to be performed during intellectual analysis, particularly in the case of tabular data. The importance of its implementation is determined by the need to reduce the sensitivity of the artificial intelligence model to the values of the features in the dataset to increase the studied model's adequacy. This paper focuses on the problem of effectively preprocessing data to improve the accuracy of intellectual analysis in the case of performing medical diagnostic tasks. We developed a new two-step method for data normalization of numerical medical datasets. It is based on the possibility of considering both the interdependencies between the features of each observation from the dataset and their absolute values to improve the accuracy when performing medical data mining tasks. We describe and substantiate each step of the algorithmic implementation of the method. We also visualize the results of the proposed method. The proposed method was modeled using six different machine learning methods based on decision trees when performing binary and multiclass classification tasks. We used six real-world, freely available medical datasets with different numbers of vectors, attributes, and classes to conduct experiments. A comparison between the effectiveness of the developed method and that of five existing data normalization methods was carried out. It was experimentally established that the developed method increases the accuracy of the Decision Tree and Extra Trees Classifier by 1–5% in the case of performing the binary classification task and the accuracy of the Bagging, Decision Tree, and Extra Trees Classifier by 1–6% in the case of performing the multiclass classification task. Increasing the accuracy of these classifiers only by using the new data normalization method satisfies all the prerequisites for its application in practice when performing various medical data mining tasks.

Keywords: medical diagnostics; classification accuracy; preprocessing; data normalization; scalars; small data; machine learning; decision trees; binary classification; multiclass classification; precision model

MSC: 15A04



Citation: Izonin, I.; Tkachenko, R.; Shakhovska, N.; Ilchyshyn, B.; Singh, K.K. A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain. *Mathematics* **2022**, *10*, 1942. <https://doi.org/10.3390/math10111942>

Academic Editor: José Antonio Sanz

Received: 11 May 2022

Accepted: 4 June 2022

Published: 6 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The rapid development of artificial intelligence tools, the widespread use of Internet of Things technologies, and the rapid growth of the computing power of modern hardware satisfy all the prerequisites for the use of intellectual analysis in various applications. It is also facilitated by the collection and preservation of large arrays of different types of data for research [1].

The data mining methodology includes three main steps: preprocessing the collected data, selecting and applying the optimal machine learning model for their analysis, and evaluating the result [2].

Data preprocessing is the first, and perhaps the most critical, step in the further analysis of such data. Effectively performing preprocessing tasks is essential to improving the accuracy of classifiers and regressors based on such data [3–5]. Numerical data preprocessing tasks include data consolidation, deduplication, data imputation, detection and removal of anomalies and outliers, feature selection, and data normalization.

In this article, we investigate the last of these data processing stages. Data normalization transforms the value of a feature in the initial dataset into a given range. The need for such a step is determined by the possible sensitivity of the selected machine learning model to the value of the feature. Thus, a non-normalized dataset can provoke a finding by the chosen machine learning model of false dependencies in the data and, as a result, reduce the efficiency of its work in performing the stated task [2,6].

Numerical data normalization is not a new problem. There are many approaches to performing it. Some methods are often used to perform this task. A number of them have been successfully implemented and used by researchers in data mining application packages. Other, more specialized methods are used in some cases. However, the general problem of selecting the optimal strategy for each specific task or dataset or machine learning model in order to obtain the highest accuracy urgently needs to be solved.

The modern development of medical diagnostics is primarily based on data mining. It happens for many reasons, such as [7]:

- The existence of historical data of different volumes intended for analysis;
- The need to analyze both enormous and tiny datasets that are difficult for humans to handle;
- A large number of features that may affect the patient's diagnosis and are difficult or impossible for doctors to take into account during diagnosis;
- Complex, usually hidden, nonlinear interdependencies between the features of a particular dataset, which are very difficult to identify at first glance but are easily identified and taken into account by a specific machine learning model;
- The high classification or prediction accuracy of machine learning models, which exclude human factors and subjectivism and can serve as a source of additional information to the doctor.

All this greatly complicates the application of medical data mining in various fields of medicine.

Despite this, the number of studies developing new and effective diagnostic technologies based on different types and volumes of information about the patient is growing every day [8,9]. All of them use a particular procedure for the normalization of the studied dataset. Selection of the optimal algorithm for or approach to data normalization can increase the performance and classification accuracy of machine learning models [10,11]. Such a simple procedure can provide a better machine learning model for medical data mining [12].

The vast majority of existing data normalization methods involve the performance of transformations on the columns of the tabular dataset. Such changes aim to reduce the value of each feature in the studied dataset to some value determined within a specific interval while maintaining the overall data distribution. This approach reduces the sensitivity and, as a result, increases the generalizability of the chosen machine learning model and can also reduce the duration of learning procedures, for example, when the values of significant features are reduced to values in a small interval (e.g., 0:1 or $-1:1$).

However, as noted above, attributes with complex, hidden, and nonlinear interdependencies characterize medical data processing tasks. These should be taken into account in the machine learning model in order to improve the accuracy of intelligent diagnostic systems. However, most of the existing methods do not yield a dataset that considers these features of medical data processing tasks.

This paper aims to develop a new data normalization method that considers the interdependencies between features in a given dataset and their absolute values. The

proposed method should increase the classification accuracy of machine learning methods in the case of medical data processing tasks.

The main contributions of this paper can be summarized as follows:

- We develop a new two-step method for tabular data normalization that considers the interdependencies between the features of each observation and the absolute values of each of these features. The proposed method reduces the number of extrapolation problems for vectors at a distance from the training sample;
- We demonstrate the high efficiency of Decision Tree and Extra Trees classifiers based on the developed data normalization method for both binary and multiclass classification tasks using different medical datasets;
- We experimentally establish an increase in the classification accuracy based on several machine learning methods that use the developed two-step data normalization method compared with other existing methods.

The remainder of the paper is structured as follows. Section 2 presents the results of a review and critical analysis of existing work on the normalization of tabular datasets. Section 3 introduces the mathematical basis of five existing data normalization methods. The developed two-step data normalization method for the medical domain is described. The algorithmic procedure for its realization and a visualization of the results are also given. Section 4 presents the numerical results of the developed method based on six different classifiers using different sets of medical data to perform binary and multiclass medical diagnostics tasks. Section 5 compares the accuracy of the developed method with that of the existing ones. Our conclusions are presented in Section 6.

2. The State-of-the-Art

Data normalization is one of the primary tasks of data processing. The performance of machine learning algorithms largely depends on how effectively the data are normalized. In particular, ref. [13] investigated the influence of different normalization methods on the accuracy of classification techniques. Based on numerous experimental studies, the author identified many techniques that provide high accuracy in classification tasks and those that should not be used to perform such tasks.

This section summarizes research on the use of several normalization methods for numerical sets of medical data and their impact on the accuracy of medical data mining techniques.

In [14], the efficiency of the k-NN classifier was investigated using different normalization methods. In particular, the authors performed experiments on the use of the Min-Max Scaler and the Standard Scaler in the selected algorithm to perform a multiclass classification task. The simulation was performed on one well-known dataset. It was experimentally established that the Min-Max Scaler provided the k-NN classifier with the highest accuracy when performing the classification task on the Iris dataset.

In [15], the results obtained in the above-mentioned study were extended. In this case, in addition to the two above-mentioned normalization methods, the authors used Decimal normalization. Moreover, the experimental part of the work analyzed the effectiveness of the application of nine machine learning methods. However, empirical studies on several datasets did not allow the authors to single out a data normalization method that would increase the accuracy of all classifiers. The authors found that the classification accuracy when using the three normalization methods varied depending on the selected classifier. The disadvantage of this study is the small number of classifiers used, which made it impossible to summarize the results on the effectiveness of a particular data normalization method.

The authors of [16] conducted experimental studies on the influence of four data normalization methods on the accuracy of an adaptive neural fuzzy inference system in performing classification and regression tasks. In addition to the three methods mentioned above, the authors used the Robust Scaler and the Max Abs Scaler. The simulation was performed using just one medical dataset. The results demonstrate that the Min-Max Scaler provided the proposed classifier with the highest accuracy when performing the medical

diagnostics task. However, experiments on only one dataset do not provide us with the possibility of generalizing the results obtained.

In [17], the authors investigated the effectiveness of performing a heart disease classification task based on different methods for normalization using nine machine learning algorithms. In particular, the authors used such normalization methods as the Robust Scaler, the Max Abs Scaler, Normalization, the Min-Max Scaler, the Standard Scaler, and the Quantile Transformer. It was experimentally established that none of the normalization methods significantly affected the effectiveness of each of the nine machine learning algorithms. There were two reasons for this. The first one is that the authors used only one dataset in their modeling. The second one is more important. The methods studied in the paper only transform data in columns. Thus, the interdependencies between the features in the studied medical dataset were not taken into account.

The authors in [18] considered five data normalization methods, including four from a previous study and the Vector Scaler. The basis of this method is that it takes into account the norm of each vector in order to normalize the dataset by rows to overcome the above-mentioned shortcoming. The authors investigated the influence of normalization methods on multi-criteria decision-making tasks. The effectiveness of each of the studied methods was evaluated using the Pearson's correlation coefficient. The authors found that the Max Abs Scaler was the most acceptable for the stated task.

In [19], the authors considered the problem of improving the classification accuracy in medical diagnostics tasks by applying an effective data normalization method. In addition to the commonly used techniques employed in the above-mentioned study, the authors drew attention to the accuracy of classifiers that use the Vector Scaler. This was due to the specific characteristics of medical diagnostics tasks, which are significantly different from those of the task performed in the above-mentioned study. Experimental results on two different datasets using three machine learning methods based on decision trees showed a significant increase in the classification accuracy in the case of using the Vector Scaler compared with the other methods. Despite this fact, such an approach does not consider the absolute values of the features in the normalized dataset. This can lead to some ambiguities that, in turn, will reduce the effectiveness of further medical data mining.

In general, most of the published scientific papers on the effect of data normalization on classification accuracy did not use methods that take into account the interdependencies between the attributes of each vector and their absolute values. However, the importance of this problem has been confirmed by many studies in various fields of biology and medicine [20–22].

In this paper, we present a new method for the normalization of numerical sets of medical data that has the advantages of the above-mentioned techniques and, at the same time, eliminates the shortcomings of these techniques in order to improve the classification accuracy of classifiers that perform medical diagnostics tasks.

3. Materials and Methods

In this paper, we present a new two-step data normalization method. It is based on the combined use of the Max Abs Scaler and the Vector Scaler, taking into account some significant differences. Therefore, we consider the principles of operation of the most common data normalization methods for numerical datasets when performing medical data mining tasks (Table 1).

The first and fourth methods are susceptible to outliers in the dataset, which is a typical characteristic of medical datasets. Additionally, if the data are not normally distributed, these are not the best Scalers to use. The Robust Scaler's centering and scaling statistics are based on percentiles and are therefore not influenced by a few large marginal outliers. The Standard Scaler assumes that the data are normally distributed within each feature, which in real-world medical datasets is impossible. Unit Vector Scaling considers the whole feature vector to be of unit length. This usually means dividing each component by the Euclidean length of the vector (i.e., using the L2 Norm).

Table 1. The most commonly used tabular data normalization methods in medical diagnostics.

#	Data Normalization Method	Mathematical Expression
1	Min Max Scaler	$x' = \frac{x_i - \min(x)}{\max(x) - \min(x)}$
2	Max Abs Scaler	$x' = \frac{x_i}{ \max(x) }$
3	Robust Scaler	$x' = \frac{x_i - \text{med}(x)}{IQR}$
4	Standard Scaler	$x' = \frac{x_i - \text{mean}(x)}{\text{std}(x)}$
5	Vector Scaler	$x' = \frac{x_i}{\sqrt{\sum_{j=1}^n (x_j)^2}}$

where x' is the normalized attribute; x_i is the current feature of the initial dataset; $\min(x)$ is the minimal value of the attribute x_i ; $\max(x)$ is the maximal value of the attribute x_i ; $\text{mean}(x)$ is the mean value of the attribute x_i ; $\text{med}(x)$ is the median value of the attribute x_i ; $\text{std}(x)$ is the standard deviation of the attribute x_i ; and IQR is the quantile range between the first and third quantiles.

In addition, the first four methods listed in Table 1 perform only column operations. Accordingly, interdependencies between the features of each vector, which are quite common in medical data, are not considered. The fifth normalization method takes into account this shortcoming. It performs normalization for each vector separately based on the norm of the corresponding vector. However, this method does not consider the absolute values of the normalized dataset.

The two-step data normalization method presented in this paper overcomes these disadvantages.

The Proposed Two-Step Data Normalization Method

The proposed data normalization method considers both the interdependencies between the features of each vector and the absolute values of each of the features in a given medical dataset. The need for this can be explained by the peculiarities of medical diagnostics tasks [23,24]. They are characterized by datasets of different volumes, with an asymmetrically represented number of vectors in each problem class. In addition, such datasets are characterized by many additional attributes (e.g., laboratory tests, physician observations) that also have complex, nonlinear, and seemingly unknown interdependencies [25]. However, considering such interdependencies is essential in diagnosis and therapy or supporting the treatment process [26]. Existing methods for normalization mainly involve the conversion of data by columns. However, this is insufficient when it is necessary to consider the interdependence between them [19]. That is why the developed method takes into account the above-mentioned features of medical datasets.

Now, we consider the developed method in more detail. Assume that a medical dataset can be represented as a matrix of features $D = [x_{i,j}]_{i=1,N}^{j=1,n}$, where each i -th vector (line, or observation) can be represented as follows: $x_i = x_{i,1}, \dots, x_{i,j}, \dots, x_{i,n}$, where $i = 1, N$ and N is the number of vectors (the number of observations in a matrix D).

The algorithmic implementation of the proposed two-step data normalization method involves the sequential execution of the following procedures.

1. Initial normalization for each j -th column ($j = 1, n$) of a given set of tabular data, according to the scheme of the maximal value of the absolute element in each column, according to the following formula:

$$x'_{i,j} = \frac{x_{i,j}}{\max_{1 \leq j \leq n} |x_{i,j}|}, \quad i = \overline{1, N}, \quad j = \overline{1, n}. \tag{1}$$

This step of the proposed method corresponds to normalization according to the second method listed in Table 1. It can be omitted or replaced by another method that normalizes the data by columns.

Accordingly, as a result of this step, we normalize the entire dataset (if it is one matrix, represented as D). If the dataset before normalization was divided into two datasets (a

training dataset and a test dataset), then the first step of the algorithm is performed on the training dataset. Next, the normalization of the test/validation dataset is completed according to the maximal value of the absolute elements for each column that were obtained for the training dataset. The same approach is used for all further steps of the proposed method in the case where the separate normalization of the training and test/validation datasets is needed.

The first step of the developed method for normalization by rows involves:

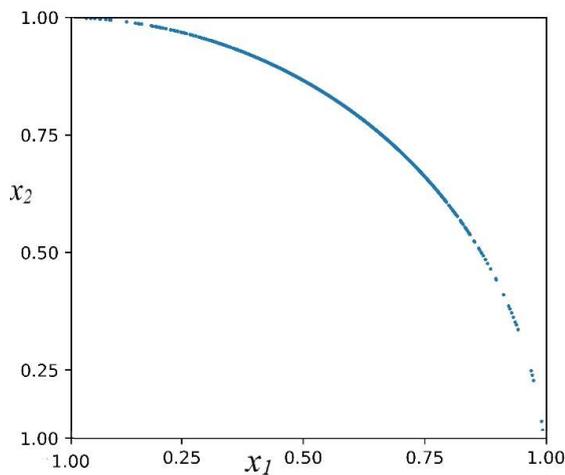
2. Calculation of the norm of each vector using $x'_{i,j}$ from (2) according to the following expression:

$$\|X'_i\| = \sqrt{\sum_{j=1}^n (x'_{i,j})^2}, \tag{2}$$

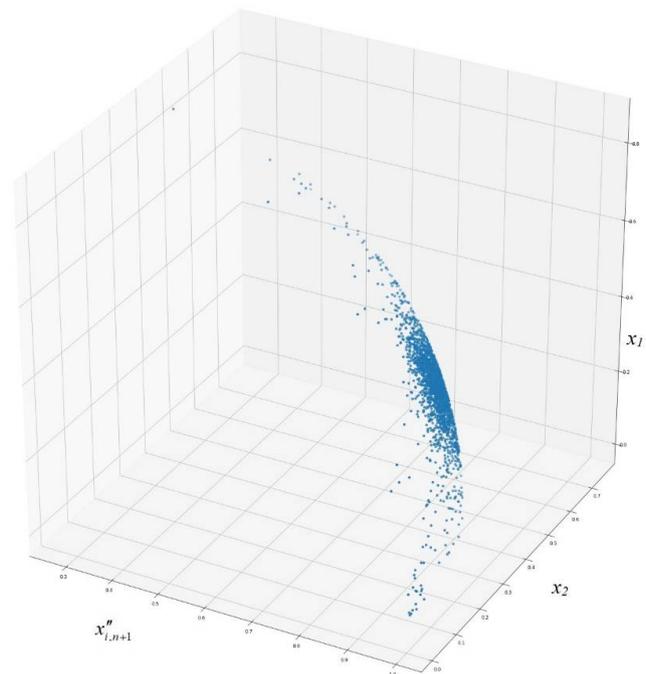
3. Normalization of each separate vector $x'_{i,j}$ from the dataset, taking into account its norm according to the expression:

$$x''_{i,j} = \frac{x'_{i,j}}{\sqrt{\sum_{j=1}^n (x'_{i,j})^2}}. \tag{3}$$

As a result, we obtain the normalization of the dataset according to Method 5 from Table 1. A visualization of the results of the proposed method for the case of a two-dimensional dataset is presented in Figure 1a.



(a)



(b)

Figure 1. Visualization of the results of two data normalization methods. (a) Vector Scaler; (b) Proposed Scaler.

The main idea is to normalize each vector (row, observation) of a given dataset separately from each other vector. The main advantage is that the normalized dataset considers the interconnections between the attributes of each observation. It is essential that this condition be satisfied in order to improve the efficiency of data mining when performing classification tasks in various fields of medicine.

However, the main disadvantage of this method is that it does not consider the absolute values of each feature. As a result, ambiguities may arise that will significantly affect the performance of the classifiers or regressors that process the dataset in this way.

We propose a second step of transformation that eliminates the above shortcoming. The second step of the proposed data normalization method transforms the data by rows.

Let us add the notation:

$$x''_{i,n+1} = \|X'_i\|. \tag{4}$$

- Then, we expand each vector (3) of the dataset using each corresponding norm (2):

$$X''_{i,j} = x''_{i,1}, \dots, x''_{i,n}, x''_{i,n+1} \tag{5}$$

As a result, we obtain a new vector with an additional input component $x''_{i,n+1}$.

- We perform for each extended vector (5) transformations similar to procedure (3):

$$x'''_{i,j} = \frac{x''_{i,j}}{\sqrt{\sum_{j=1}^{n+1} (x''_{i,j})^2}} \tag{6}$$

In this case, we calculate the norm $\|X''_i\|$ of each extended vector from (5) and normalize each vector for the second time taking into account its new norm.

A visualization of the results of the proposed two-step data normalization method for the case of an initial two-dimensional dataset ($x_1, x_2 \rightarrow y$) is presented in Figure 1b.

As a result of this step, we obtain:

- A normalized dataset for each column and each row;
- A dataset that has been extended by one additional feature compared with the original, non-normalized dataset;
- A dataset that considers both the interdependencies between the features of each separate vector and their absolute values.

If we analyze the results of both normalization methods for the case of the initial two-dimensional dataset, we can obtain the following conclusions. The result of the Vector Scaler normalization method (Figure 1a) is a set of vectors that lie on a circle of unit radius. This method allows for the interdependence between the attributes of a given dataset to be considered but not their absolute value. If we use the proposed two-step data normalization method on rows (Figure 1b), the obtained set of normalized vectors will lie on a sphere. This is due to the introduction of an additional component in each vector of the two-dimensional data array; therefore, the visualization occurs in three-dimensional space. In this case, the third component considers the absolute values of the vectors. For example, using two vectors with components (5, 6) and (10, 12) will ensure the possibility of distinguishing them in the normalized dataset. According to the Vector Scaler, the normalized components of both these vectors will be the same. This reduces the informativeness of the whole dataset. In the case of small data, processing them can be a problem. The proposed two-step data normalization method increases the dimensionality of the input data space by adding a third component that considers the absolute values of the vector components. This ensures that the selected classifier will be able to separate these two vectors.

Among the apparent consequences of implementing the proposed approach is that the projection on a sphere will reduce the number of extrapolation problems for vectors at a distance from the training sample. Therefore, applying the proposed two-step data normalization method should increase the classifier's accuracy when performing various medical diagnostics tasks.

4. Modeling and Results

We developed a software solution to implement the two-step data normalization method using Python [27]. The simulation of the proposed method was performed using several machine learning methods based on decision trees. We used two boosting machine learning methods, bagging and feature bagging methods, and Decision and Extra Precision Trees methods [28]. This choice was due to their high accuracy, the possibility of straightforwardly interpreting some of their results, and the widespread use of such methods to perform various technical and medical diagnostics tasks [29,30].

The modeling used fixed parameters for each of the machine learning methods used for each of the studied data normalization methods. To easily reproduce the results of this study, we chose the implementation of each of the machine learning methods in the Python library, namely Scikit-learn [16]. The parameters of the methods used during the modeling are summarized in Table A1.

The evaluation of the accuracy of the machine learning methods with the proposed data normalization method and the existing data normalization methods was carried out using standard performance indicators. In particular, Accuracy, Precision, Recall, and F1-score were used to assess the effectiveness of the classifiers in performing the tasks [31,32].

4.1. Datasets Used for the Modeling

We investigated whether the proposed data normalization method increases the accuracy of classifiers that perform medical diagnostics tasks [33]. Most of these tasks are formulated as classification tasks with two or more defined classes [34,35]. If there are only two classes, which is very common in medical diagnostics tasks, we consider a binary classification task [36]. If the problem has more than two defined classes, it is a multiclass classification task [37].

Since both formulations are typical of applied medical diagnostics tasks, we modeled the proposed two-step data normalization method on different medical datasets designed to perform binary and multiclass classification tasks. To do this, we selected three well-known, real-world datasets for the binary classification task and three real-world, well-known datasets for the multiclass classification task. It should be noted that the number of data vectors in each dataset and the number of features in each dataset are different. In addition, the datasets for the multiclass classification task had between three and six classes.

A summary of the datasets used for the modeling and references to the freely available repository where they are located are given in Table 2.

Table 2. Datasets used for the modeling and their main characteristics.

Dataset Title	Problem	Attributes	Vectors	Classes	Reference
Heart Attack Analysis & Prediction Dataset	Binary classification	13	303	2	[38]
Blood Transfusion Service Center Dataset	Binary classification	4	748	2	[39]
Heart Failure Prediction Dataset	Binary classification	12	299	2	[40]
Maternal Health Risk Dataset	Multiclass classification	6	1014	3	[41]
Breast Tissue Dataset	Multiclass classification	9	212	6	[42]
Contraceptive Method Choice Dataset	Multiclass classification	9	1473	3	[43]

Each dataset was divided into two datasets: a training dataset (80% of the samples, randomly selected) and a test dataset (the remaining 20% of the samples).

4.2. Results

Table 3 summarizes the results of modeling the proposed two-step data normalization method based on the:

1. Accuracy score;
2. Precision score;
3. Recall score; and
4. F1-score

using the six different machine learning methods for the six datasets.

Table 3. Values of the four performance indicators for the classification accuracy of the proposed data normalization method based on the six machine learning models using the six different datasets.

Dataset Title	Classifier	Accuracy Score	Precision Score *	Recall Score *	F1-Score *
Heart Attack Analysis & Prediction Dataset	AdaBoost	0.770	0.844	0.750	0.794
	Bagging	0.787	0.795	0.861	0.827
	Decision Tree	0.754	0.784	0.806	0.795
	Extra Trees	0.803	0.800	0.889	0.842
	Gradient Boosting	0.770	0.806	0.806	0.806
	Random Forest	0.787	0.811	0.833	0.822
Blood Transfusion Service Center Dataset	AdaBoost	0.707	0.731	0.925	0.817
	Bagging	0.713	0.756	0.877	0.812
	Decision Tree	0.680	0.742	0.840	0.788
	Extra Trees	0.740	0.768	0.906	0.831
	Gradient Boosting	0.707	0.746	0.887	0.810
	Random Forest	0.720	0.767	0.868	0.814
Heart Failure Prediction Dataset	AdaBoost	0.800	0.667	0.588	0.625
	Bagging	0.767	0.615	0.471	0.533
	Decision Tree	0.750	0.556	0.588	0.571
	Extra Trees	0.733	0.625	0.294	0.4
	Gradient Boosting	0.800	0.692	0.529	0.600
	Random Forest	0.817	0.750	0.529	0.621
Maternal Health Risk Dataset	AdaBoost	0.562	0.589	0.562	0.563
	Bagging	0.852	0.854	0.852	0.852
	Decision Tree	0.857	0.859	0.857	0.857
	Extra Trees	0.857	0.858	0.857	0.857
	Gradient Boosting	0.783	0.785	0.783	0.783
	Random Forest	0.837	0.840	0.837	0.837
Breast Tissue Dataset	AdaBoost	0.429	0.313	0.429	0.324
	Bagging	0.667	0.829	0.667	0.683
	Decision Tree	0.571	0.786	0.571	0.638
	Extra Trees	0.667	0.749	0.667	0.695
	Gradient Boosting	0.571	0.683	0.571	0.598
	Random Forest	0.571	0.698	0.571	0.591
Contraceptive Method Choice Dataset	AdaBoost	0.475	0.482	0.475	0.476
	Bagging	0.508	0.510	0.508	0.508
	Decision Tree	0.451	0.453	0.451	0.452
	Extra Trees	0.485	0.483	0.485	0.482
	Gradient Boosting	0.539	0.541	0.539	0.537
	Random Forest	0.502	0.503	0.502	0.501

* It should be noted that the Precision, Recall, and F1-scores were calculated by finding their average, weighted by the support.

As shown in Table 3, the Extra Decision Tree classifier had the highest accuracy among all the methods considered in most cases. Additionally, in some cases, the Bagging classifier and the Decision Tree classifier showed a good trend. In contrast, the ensemble techniques, the AdaBoost Classifier and the Random Forest Classifier, had low classification accuracy in the stated tasks.

5. Comparison and Discussion

The effectiveness of the proposed data normalization method was evaluated by comparing its accuracy with that of five existing data normalization methods:

- Vector Scaler;
- Max Abs Scaler;
- Min Max Scaler;

- Standard Scaler;
- Robust Scaler.

The results (the Accuracy score and the F1-score) of the six machine learning methods on the six datasets using the six data normalization methods are summarized in Tables A2 and A3.

Since we considered binary and multiclass classification tasks, the analysis of the results for each was carried out separately.

In performing the binary classification task, some of the machine learning methods, namely the AdaBoost classifier, the Bagging classifier, the Gradient Boosting classifier, and the Random Forest classifier, demonstrated a deterioration in accuracy when using the proposed two-step data normalization method.

The best effect when using the proposed method to perform the binary classification task was obtained by using the two most straightforward machine learning methods (the Decision Tree classifier and the Extra Trees classifier), the results of which are easy to interpret. This advantage is essential when performing medical data mining, where the latest trend is the use of Explainable Artificial Intelligence.

Figure 2 summarizes the results of the use of both these methods (based on the Accuracy score) to perform binary classification tasks using three different datasets and six different data normalization methods.

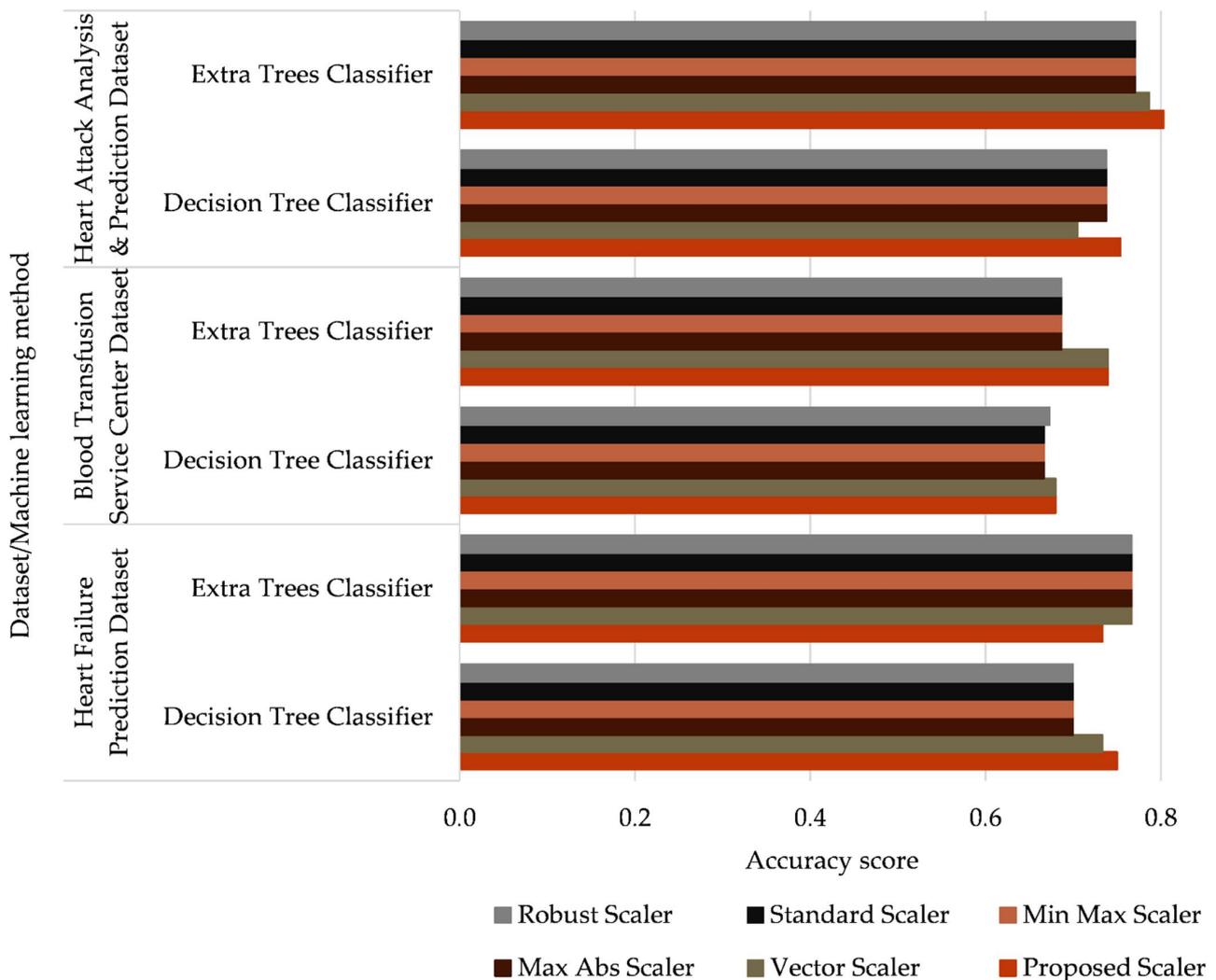


Figure 2. Accuracy scores for two machine learning methods used to perform binary classification tasks on three medical datasets using six data normalization methods.

It should be noted that the numerical values of the Accuracy score accompany the graphical information in two of the six columns. These are the values that were obtained for the proposed method and the most similar method. For the other methods, these numbers are not given so as to not overload the histogram. All accuracy indicators are presented in Table A1.

Figure 2 summarizes the accuracy of both methods during the completion of the binary classification task using the three different datasets and the six data normalization methods.

As can be seen from Figure 2, when analyzing the dataset [38] produced by the Extra Trees classifier, the proposed method showed a 1% increase in accuracy compared with the most similar method (the Vector Scaler) and a 2% increase in accuracy compared with all other normalization methods. Using the Decision Tree classifier, we obtained a 5% increase in accuracy compared with the most similar classifier and a 3% increase in accuracy compared with all other normalization methods.

In the case of the analysis of the dataset from [39], the Extra Trees classifier based on the proposed method demonstrated a 5% increase in accuracy compared with all other methods. The Decision Tree classifier, in this case, showed an increase in accuracy of more than 1%.

In the third case, when completing a binary classification task based on the dataset from [40], the Extra Trees classifier based on the proposed method experienced a 2% reduction in accuracy. The Decision Tree classifier showed a 2% increase in accuracy compared with all other methods.

Taken together, the results of the Decision Tree classifier and the Extra Trees classifier based on the six different data normalization methods for the binary classification task indicate that:

- The Max Abs Scaler, Min-Max Scaler, Standard Scaler, and Robust Scaler do not provide a significant difference in the accuracy of the investigated classifiers;
- The proposed data normalization method provides an increase in the classification accuracy of 1 to 5% compared with the existing methods;
- The proposed data normalization method increases the classification accuracy from 1% to 3% compared with the most similar data normalization method (the Vector Scaler).

Let us now consider the results of the comparison between the proposed method and the existing normalization methods in the case of completing multiclass classification tasks based on the three studied datasets and the six different machine learning methods. The results are presented in Table A2.

In performing multiclass classification tasks, some machine learning methods, namely the AdaBoost classifier, the Gradient Boosting classifier, and the Random Forest classifier, demonstrated a deterioration in accuracy when using the proposed two-step data normalization method. An increase in accuracy with the proposed method, in this case, was achieved using the Bagging classifier, the Decision Tree classifier, and the Extra Trees classifier.

Figure 3 summarizes the accuracy of these methods based on the F1-score during the completion of multiclass classification tasks using the three different datasets and the six different data normalization methods.

As shown in Figure 3, when completing a multiclass classification task based on the dataset from [41], four well-known data normalization methods had almost no effect on the accuracy of the classifiers. Here again, the proposed method and the most similar method stand out. For this dataset (the dataset from [41]), all three machine learning methods experienced a 1% to 6% increase in accuracy due to the proposed data normalization method. Similar results were obtained for the dataset from [42]. Here, the Bagging classifier and the Extra Trees classifier demonstrated a significant increase in classification accuracy. This can be explained by the fact that the proposed method increases the number of features in the dataset by one, which results in the jump in the accuracy of the classifiers. However, in the case of using the dataset from [43], the accuracy of the classifiers using all six data

normalization methods is almost the same. Only the Bagging classifier showed an increase in accuracy (1%) when using the proposed method.

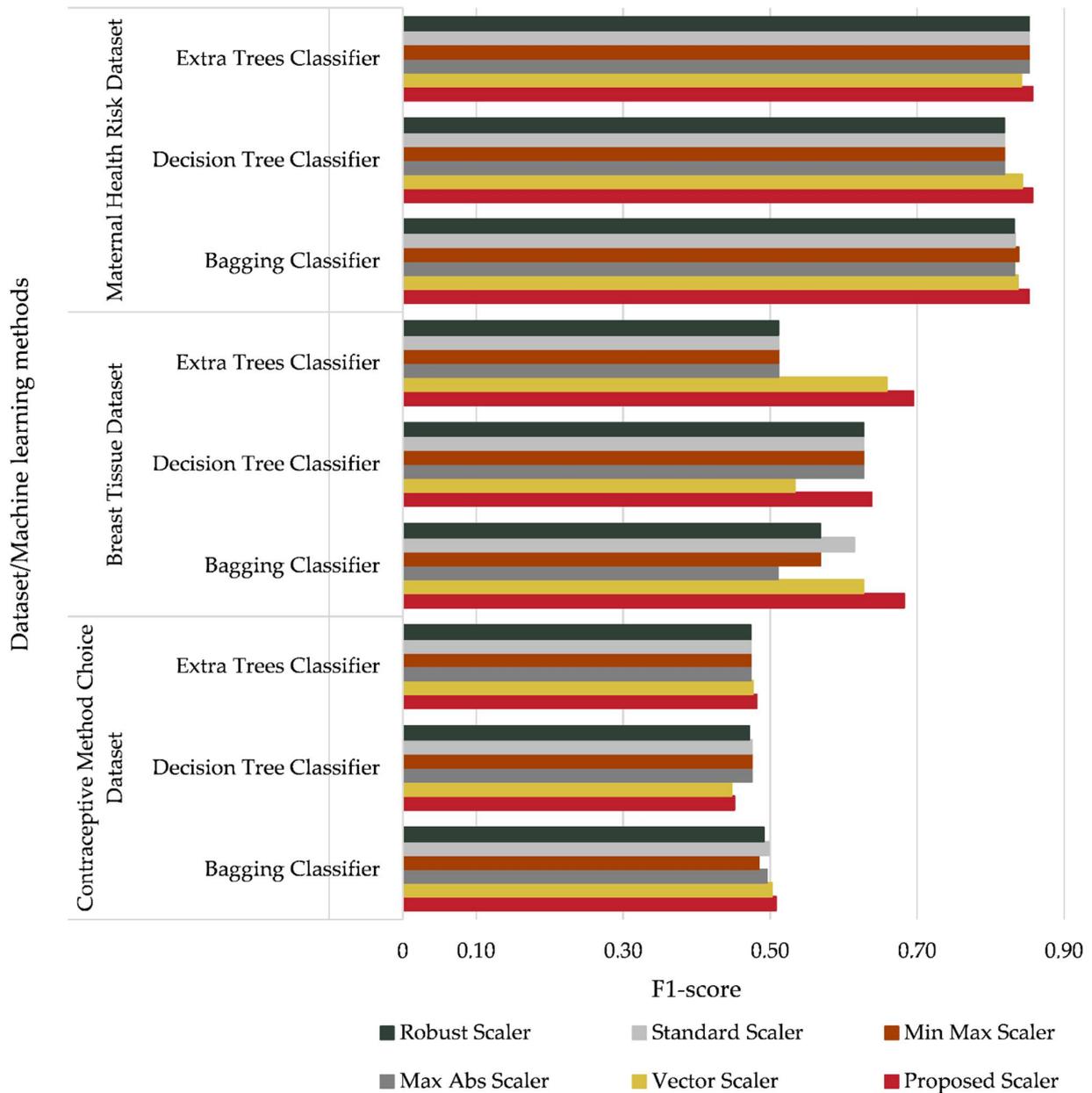


Figure 3. F1-scores for three machine learning methods used to perform multiclass classification tasks on three medical datasets using six data normalization methods.

It should be noted that all variables in the third dataset are categorical, which may explain the generally low accuracy of the machine learning methods applied for its analysis.

Taken together, the results of the Bagging classifier, the Decision Tree classifier, and the Extra Trees classifier based on six different data normalization methods for multiclass classification tasks indicate that:

- The Max Abs Scaler, Min-Max Scaler, Standard Scaler, and Robust Scaler affect the accuracy of the investigated classifiers;
- The proposed data normalization method provides both a significant (1% to 6%) increase in the accuracy of the classifiers compared with the above-mentioned methods for normalization and the same level of accuracy as the Vector Scaler;

- The proposed data normalization method improves the accuracy of the classifiers compared with the most similar data normalization method (the Vector Scaler).

In general, an increase in the accuracy of a classifier of 1% based on only the data normalization method, which is perhaps the first step in data mining, would justify its use in practice. However, increasing the accuracy by 5% in binary classification tasks only by normalizing the data satisfies many prerequisites for using the proposed method in Decision Tree and Extra Trees classifiers that perform various medical diagnostics tasks, particularly in automated robotic systems [44–46]. Such a significant increase in the accuracy of the Bagging classifier, the Decision Tree classifier, and the Extra Trees classifier based on the proposed data normalization method in performing multiclass classification tasks also encourages the use of the Extra Trees classifier in practice.

6. Conclusions

In this study, we focused on the problem of effectively preprocessing data to increase the accuracy of intellectual analysis in the case of completing medical diagnostics tasks. We developed a new two-step numerical data normalization method. It is based on the possibility of considering the interdependencies between the features of each observation and the absolute values of each of these features to improve the accuracy of medical data mining techniques.

The proposed approach was modeled using six different classifiers based on machine learning methods for the two cases of binary classification tasks and multiclass classification tasks. Experiments were performed on six real-world, freely available datasets for performing medical diagnostics tasks with different numbers of vectors, attributes, and classes.

We compared the accuracy of the proposed data normalization method with that of five existing methods. It was established that the proposed data normalization method increased the classification accuracy of the Decision Tree classifier and the Extra Trees classifier by 1–5% in the case of performing the binary classification task. In addition, it provided a 1–6% increase in the accuracy of the Bagging classifier, the Decision Tree classifier, and the Extra Trees classifier in the case of performing the multiclass classification task. At the same time, we observed a decrease in the classification accuracy of the AdaBoost classifier, the Gradient Boosting classifier, and the Random Forest classifier when using the proposed normalization method compared with the existing ones in both classification tasks.

Nevertheless, the increase in the accuracy of the Decision Tree classifier and the Extra Trees classifier based only on the proposed data normalization method satisfies all the prerequisites for its use in practice when performing a variety of medical data mining tasks.

Further research will be conducted to assess the accuracy of artificial neural networks [46–48], particularly PNN and GRNN, based on the developed two-step data normalization method for the analysis of small datasets. In addition, using the method proposed in this paper, a new data classification method will be developed for imbalanced datasets and the representation of only the vectors of one class in the dataset, which should be recognized given that many vectors have previously not been described. This method will be based on the new committee model of a hypercylinder's surfaces based on nonlinear SGTm neural-like structures [49].

Author Contributions: Conceptualization, R.T. and I.I.; methodology, I.I.; software, B.I.; validation, R.T., K.K.S. and N.S.; formal analysis, R.T.; investigation, I.I.; resources, N.S.; data curation, B.I. and K.K.S.; writing—original draft preparation, I.I.; writing—review and editing, I.I. and N.S.; visualization, N.S.; supervision, R.T.; project administration, I.I. and K.K.S.; funding acquisition, N.S. and I.I. All authors have read and agreed to the published version of the manuscript.

Funding: The National Research Foundation of Ukraine funded this study under project number 2021.01/0103.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data presented in this study are available in publicly accessible repositories, links to which can be found in [38–43].

Acknowledgments: The authors would like to thank the anonymous reviewers for their concise recommendations that helped us present the materials better. We would also like to thank the Armed Forces of Ukraine for providing the security required to perform this work. This work was possible only because of the resilience and courage of the Ukrainian Army.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Parameters of the investigated ML-based classifiers.

ML-Based Classifier	Parameters
AdaBoost Classifier	base_estimator = None, n_estimators = 100, learning_rate = 1.0, algorithm = ‘SAMME.R’, random_state = None
Bagging Classifier	base_estimator = None, n_estimators = 100, max_samples = 1.0, max_features = 1.0, bootstrap = True, bootstrap_features = False, oob_score = False, warm_start = False, n_jobs = None, random_state = None, verbose = 0
Decision Tree Classifier	max_depth = None, min_samples_split = 2, random_state = 0
Extra Trees Classifier	n_estimators = 100, max_depth = None, min_samples_split = 2, random_state = 0
Gradient Boosting Classifier	loss = ‘deviance’, learning_rate = 0.1, n_estimators = 100, subsample = 1.0, criterion = ‘friedman_mse’, min_samples_split = 2, min_samples_leaf = 1, min_weight_fraction_leaf = 0.0, max_depth = 3, min_impurity_decrease = 0.0, init = None, random_state = None, max_features = None, verbose = 0, max_leaf_nodes = None, warm_start = False, validation_fraction = 0.1, n_iter_no_change = None, tol = 0.0001, ccp_alpha = 0.0
Random Forest Classifier	n_estimators = 100, max_depth = None, min_samples_split = 2, random_state = 0

Table A2. Accuracy scores for the six machine learning models using the six data normalization methods.

Dataset Title	Classifier	Proposed Scaler	Vector Scaler	Max Abs Scaler	Min Max Scaler	Standard Scaler	Robust Scaler
Heart Attack Analysis & Prediction Dataset	AdaBoost	0.770	0.770	0.787	0.787	0.770	0.787
	Bagging	0.787	0.787	0.852	0.836	0.836	0.852
	Decision Tree	0.754	0.705	0.738	0.738	0.738	0.738
	Extra Trees	0.803	0.787	0.770	0.770	0.770	0.770
	Gradient Boosting	0.770	0.770	0.787	0.787	0.787	0.787
Blood Transfusion Service Center Dataset	Random Forest	0.787	0.803	0.820	0.820	0.820	0.836
	AdaBoost	0.707	0.707	0.733	0.733	0.727	0.733
	Bagging	0.713	0.700	0.700	0.713	0.713	0.720
	Decision Tree	0.680	0.680	0.667	0.667	0.667	0.673
	Extra Trees	0.740	0.740	0.687	0.687	0.687	0.687
Heart Failure Prediction Dataset	Gradient Boosting	0.707	0.707	0.753	0.753	0.753	0.753
	Random Forest	0.720	0.720	0.713	0.707	0.720	0.720
	AdaBoost	0.800	0.800	0.750	0.750	0.750	0.750
	Bagging	0.767	0.833	0.767	0.767	0.800	0.783
	Decision Tree	0.750	0.733	0.700	0.700	0.700	0.700
Heart Failure Prediction Dataset	Extra Trees	0.733	0.767	0.767	0.767	0.767	0.767
	Gradient Boosting	0.800	0.800	0.800	0.800	0.800	0.800
	Random Forest	0.817	0.833	0.800	0.800	0.817	0.817

Table A2. *Cont.*

Dataset Title	Classifier	Proposed Scaler	Vector Scaler	Max Abs Scaler	Min Max Scaler	Standard Scaler	Robust Scaler
Maternal Health Risk Dataset	AdaBoost	0.562	0.562	0.690	0.690	0.690	0.690
	Bagging	0.852	0.837	0.833	0.837	0.833	0.833
	Decision Tree	0.857	0.842	0.818	0.818	0.818	0.818
	Extra Trees	0.857	0.842	0.852	0.852	0.852	0.852
	Gradient Boosting	0.783	0.788	0.783	0.783	0.783	0.783
	Random Forest	0.837	0.828	0.833	0.837	0.842	0.833
Breast Tissue Dataset	AdaBoost	0.429	0.429	0.524	0.524	0.524	0.524
	Bagging	0.667	0.619	0.524	0.571	0.619	0.571
	Decision Tree	0.571	0.476	0.619	0.619	0.619	0.619
	Extra Trees	0.667	0.619	0.476	0.476	0.476	0.476
	Gradient Boosting	0.571	0.571	0.619	0.619	0.619	0.619
	Random Forest	0.571	0.667	0.571	0.571	0.571	0.571
Contraceptive Method Choice Dataset	AdaBoost	0.475	0.475	0.512	0.512	0.512	0.512
	Bagging	0.508	0.502	0.495	0.485	0.498	0.492
	Decision Tree	0.451	0.447	0.475	0.475	0.475	0.471
	Extra Trees	0.485	0.478	0.475	0.475	0.475	0.475
	Gradient Boosting	0.539	0.539	0.532	0.532	0.532	0.532
	Random Forest	0.502	0.515	0.502	0.495	0.498	0.508

Table A3. F1-scores for the six machine learning models using the six data normalization methods.

Dataset Title	Classifier	Proposed Scaler	Vector Scaler	Max Abs Scaler	Min Max Scaler	Standard Scaler	Robust Scaler
Heart Attack Analysis & Prediction Dataset	AdaBoost	0.794	0.794	0.817	0.817	0.806	0.817
	Bagging	0.827	0.827	0.883	0.868	0.865	0.880
	Decision Tree	0.795	0.757	0.778	0.778	0.778	0.778
	Extra Trees	0.842	0.827	0.800	0.800	0.800	0.800
	Gradient Boosting	0.806	0.806	0.827	0.827	0.827	0.827
	Random Forest	0.822	0.838	0.849	0.849	0.849	0.865
Blood Transfusion Service Center Dataset	AdaBoost	0.817	0.817	0.831	0.831	0.826	0.831
	Bagging	0.812	0.800	0.805	0.812	0.809	0.817
	Decision Tree	0.788	0.784	0.779	0.779	0.779	0.784
	Extra Trees	0.831	0.831	0.797	0.797	0.797	0.797
	Gradient Boosting	0.810	0.810	0.843	0.843	0.843	0.843
	Random Forest	0.814	0.814	0.811	0.805	0.816	0.816
Heart Failure Prediction Dataset	AdaBoost	0.625	0.625	0.595	0.595	0.595	0.595
	Bagging	0.533	0.667	0.533	0.533	0.625	0.581
	Decision Tree	0.571	0.556	0.400	0.400	0.400	0.400
	Extra Trees	0.4	0.462	0.462	0.462	0.462	0.462
	Gradient Boosting	0.600	0.600	0.625	0.625	0.625	0.625
	Random Forest	0.621	0.667	0.625	0.625	0.645	0.645
Maternal Health Risk Dataset	AdaBoost	0.563	0.563	0.692	0.692	0.692	0.692
	Bagging	0.852	0.837	0.833	0.838	0.833	0.832
	Decision Tree	0.857	0.843	0.819	0.819	0.819	0.819
	Extra Trees	0.857	0.842	0.853	0.853	0.853	0.853
	Gradient Boosting	0.783	0.788	0.783	0.783	0.783	0.783
	Random Forest	0.837	0.828	0.833	0.838	0.843	0.833
Breast Tissue Dataset	AdaBoost	0.324	0.376	0.475	0.475	0.475	0.475
	Bagging	0.683	0.627	0.511	0.568	0.615	0.568
	Decision Tree	0.638	0.534	0.628	0.628	0.628	0.628
	Extra Trees	0.695	0.659	0.511	0.511	0.511	0.511
	Gradient Boosting	0.598	0.607	0.618	0.618	0.618	0.635
	Random Forest	0.591	0.678	0.568	0.568	0.568	0.568

Table A3. Cont.

Dataset Title	Classifier	Proposed Scaler	Vector Scaler	Max Abs Scaler	Min Max Scaler	Standard Scaler	Robust Scaler
Contraceptive Method Choice Dataset	AdaBoost	0.476	0.476	0.510	0.510	0.510	0.510
	Bagging	0.508	0.502	0.496	0.484	0.498	0.492
	Decision Tree	0.452	0.447	0.475	0.475	0.475	0.471
	Extra Trees	0.482	0.476	0.474	0.474	0.474	0.474
	Gradient Boosting	0.537	0.537	0.534	0.534	0.534	0.534
	Random Forest	0.501	0.514	0.502	0.494	0.499	0.509

References

- Kumar, P.; Kumar, Y.; Tawhid, M.A. (Eds.) *Machine Learning, Big Data, and IoT for Medical Informatics; Intelligent Data Centric Systems*; Academic Press: Cambridge, MA, USA, 2021; ISBN 978-0-12-821777-1.
- Hu, Z.; Tereykovski, I.A.; Tereykovska, L.O.; Pogorelov, V.V. Determination of Structural Parameters of Multilayer Perceptron Designed to Estimate Parameters of Technical Systems. *IJISA* **2017**, *9*, 57–62. [[CrossRef](#)]
- Shakhovska, N.; Yakovyna, V.; Kryvinska, N. An Improved Software Defect Prediction Algorithm Using Self-Organizing Maps Combined with Hierarchical Clustering and Data Preprocessing. In *International Conference on Database and Expert Systems Applications*; Hartmann, S., Küng, J., Kotsis, G., Tjoa, A.M., Khalil, I., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 414–424.
- Hu, Z.; Ivashchenko, M.; Lyushenko, L.; Klyushnyk, D. Artificial Neural Network Training Criterion Formulation Using Error Continuous Domain. *IJMECS* **2021**, *13*, 13–22. [[CrossRef](#)]
- Tlebaldinova, A.; Denissova, N.; Baklanova, O.; Krak, I.; Györök, G. Normalization of Vehicle License Plate Images Based on Analyzing of Its Specific Features for Improving the Quality Recognition. *Acta Polytech. Hung.* **2020**, *17*, 193–206. [[CrossRef](#)]
- Hu, Z.; Bodyanskiy, Y.V.; Kulishova, N.Y.; Tyshchenko, O.K. A Multidimensional Extended Neo-Fuzzy Neuron for Facial Expression Recognition. *IJISA* **2017**, *9*, 29–36. [[CrossRef](#)]
- Izonin, I.; Tkachenko, R. Universal Intraensemble Method Using Nonlinear AI Techniques for Regression Modeling of Small Medical Data Sets. In *Cognitive and Soft Computing Techniques for the Analysis of Healthcare Data*; Elsevier: Amsterdam, The Netherlands, 2022; pp. 123–150. ISBN 978-0-323-85751-2.
- Krak, I.; Barmak, O.; Manziuk, E. Using Visual Analytics to Develop Human and Machine-centric Models: A Review of Approaches and Proposed Information Technology. *Comput. Intell.* **2020**, 1–26. [[CrossRef](#)]
- Krak, Y.V. Dynamics of Manipulation Robots: Numerical-Analytical Method of Formation and Investigation of Computational Complexity. *J. Automat. Inf. Sci.* **1999**, *31*, 121–128. [[CrossRef](#)]
- Babichev, S.; Lytvynenko, V.; Škvor, J.; Korobchynskiy, M.; Voronenko, M. Information Technology of Gene Expression Profiles Processing for Purpose of Gene Regulatory Networks Reconstruction. In *Proceedings of the 2018 IEEE Second International Conference on Data Stream Mining Processing (DSMP)*, Lviv, Ukraine, 21–25 August 2018; pp. 336–341.
- Lytvynenko, V.; Wojcik, W.; Fefelov, A.; Lurie, I.; Savina, N.; Voronenko, M.; Boskin, O.; Smailova, S. Hybrid Methods of GMDH-Neural Networks Synthesis and Training for Solving Problems of Time Series Forecasting. In *Lecture Notes in Computational Intelligence and Decision Making*; Lytvynenko, V., Babichev, S., Wójcik, W., Vynokurova, O., Vyshemyrskaya, S., Radetskaya, S., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2020; Volume 1020, pp. 513–531, ISBN 978-3-030-26473-4.
- Hassler, A.P.; Menasalvas, E.; García-García, F.J.; Rodríguez-Mañas, L.; Holzinger, A. Importance of Medical Data Preprocessing in Predictive Modeling and Risk Factor Discovery for the Frailty Syndrome. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 33. [[CrossRef](#)] [[PubMed](#)]
- Singh, D.; Singh, B. Investigating the Impact of Data Normalization on Classification Performance. *Appl. Soft Comput.* **2020**, *97*, 105524. [[CrossRef](#)]
- Pandey, A.; Jain, A. Comparative Analysis of KNN Algorithm Using Various Normalization Techniques. *IJCNIS* **2017**, *9*, 36–42. [[CrossRef](#)]
- Alshdaifat, E.; Alshdaifat, D.; Alsarhan, A.; Hussein, F.; El-Salhi, S.M.F.S. The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance. *Data* **2021**, *6*, 11. [[CrossRef](#)]
- Polatgil, Mesut. Investigation of the Effect of Normalization Methods on ANFIS Success: Forestfire and Diabets Datasets. *IJITCS* **2022**, *14*, 1–8. [[CrossRef](#)]
- Ahsan, M.M.; Mahmud, M.A.P.; Saha, P.K.; Gupta, K.D.; Siddique, Z. Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. *Technologies* **2021**, *9*, 52. [[CrossRef](#)]
- Vafaei, N.; Ribeiro, R.A.; Camarinha-Matos, L.M. Normalization Techniques for Multi-Criteria Decision Making: Analytical Hierarchy Process Case Study. In *Technological Innovation for Cyber-Physical Systems*; Camarinha-Matos, L.M., Falcão, A.J., Vafaei, N., Najdi, S., Eds.; IFIP Advances in Information and Communication Technology; Springer International Publishing: Cham, Switzerland, 2016; Volume 470, pp. 261–269, ISBN 978-3-319-31164-7.

19. Izonin, I.; Tkachenko, R.; Shakhovska, N.; Ilchyshyn, B.; Gregus, M.; Strauss, C. Towards Data Normalization Task for the Efficient Mining of Medical Data. In Proceedings of the 2022 12th International Conference on Advanced Computer Information Technologies, Spišská Kapitula, Slovakia, 26–28 September 2022; pp. 1–5.
20. Nam, S.L.; de la Mata, A.P.; Dias, R.P.; Harynuk, J.J. Towards Standardization of Data Normalization Strategies to Improve Urinary Metabolomics Studies by GC×GC-TOFMS. *Metabolites* **2020**, *10*, 376. [[CrossRef](#)]
21. Viallon, V.; His, M.; Rinaldi, S.; Breeur, M.; Gicquiau, A.; Hemon, B.; Overvad, K.; Tjønneland, A.; Rostgaard-Hansen, A.L.; Rothwell, J.A.; et al. A New Pipeline for the Normalization and Pooling of Metabolomics Data. *Metabolites* **2021**, *11*, 631. [[CrossRef](#)]
22. Isaksson, F.; Lundy, L.; Hedström, A.; Székely, A.J.; Mohamed, N. Evaluating the Use of Alternative Normalization Approaches on SARS-CoV-2 Concentrations in Wastewater: Experiences from Two Catchments in Northern Sweden. *Environments* **2022**, *9*, 39. [[CrossRef](#)]
23. Chumachenko, D.; Sokolov, O.; Yakovlev, S. Fuzzy Recurrent Mappings in Multiagent Simulation of Population Dynamics Systems. *IJC* **2020**, *19*, 290–297. [[CrossRef](#)]
24. Strontsitska, A.-O.; Pavliuk, O.; Dunaev, R.; Derkachuk, R. Forecast of the Number of New Patients and Those Who Died from COVID-19 in Bahrain. In Proceedings of the 2020 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 8 November 2020; pp. 422–426.
25. Mochurad, L.; Hladun, Y. Modeling of Psychomotor Reactions of a Person Based on Modification of the Tapping Test. *Int. J. Comput.* **2021**, *20*, 1–10, in press. [[CrossRef](#)]
26. Pavliuk, O.; Strontsitska, A.-O. Combined Machine Learning Model for Covid-19 Analysis and Forecasting in Ukraine. In *The International Conference on Artificial Intelligence and Logistics Engineering*; Hu, Z., Zhang, Q., Petoukhov, S., He, M., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 16–26.
27. Hovorushchenko, T.; Pavlova, O. Method of Activity of Ontology-Based Intelligent Agent for Evaluating Initial Stages of the Software Lifecycle. In *Recent Developments in Data Science and Intelligent Analysis of Information*; Chertov, O., Mylovanov, T., Kondratenko, Y., Kacprzyk, J., Kreinovich, V., Stefanuk, V., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing: Cham, Switzerland, 2019; Volume 836, pp. 169–178, ISBN 978-3-319-97884-0.
28. API Reference. Available online: <https://scikit-learn/stable/modules/classes.html> (accessed on 8 May 2022).
29. Babenko, V.; Panchyshyn, A.; Zomchak, L.; Nehrey, M.; Artym-Drohomyretska, Z.; Lahotskyi, T. Classical Machine Learning Methods in Economics Research: Macro and Micro Level Examples. *Wseas Trans. Bus. Econ.* **2021**, *18*, 209–217. [[CrossRef](#)]
30. Rabcan, J.; Levashenko, V.; Zaitseva, E.; Kvassay, M.; Subbotin, S. Application of Fuzzy Decision Tree for Signal Classification. *IEEE Trans. Ind. Inf.* **2019**, *15*, 5425–5434. [[CrossRef](#)]
31. Rawat, B.; Dwivedi, S.K. Selecting Appropriate Metrics for Evaluation of Recommender Systems. *IJITCS* **2019**, *11*, 14–23. [[CrossRef](#)]
32. Aamir, M.; Rahman, Z.; Ahmed Abro, W.; Tahir, M.; Mustajar Ahmed, S. An Optimized Architecture of Image Classification Using Convolutional Neural Network. *IJIGSP* **2019**, *11*, 30–39. [[CrossRef](#)]
33. Khavalko, V.; Tsmots, I.; Kostyniuk, A.; Strauss, C. Classification and Recognition of Medical Images Based on the SGTm Neuroparadigm. In Proceedings of the 2nd International Workshop on Informatics & Data-Driven Medicine (IDDM 2019), Lviv, Ukraine, 11–13 November 2019; Volume 2488, pp. 234–245.
34. Bodyanskiy, Y.; Vynokurova, O.; Savvo, V.; Tverdokhlib, T.; Mulesa, P. Hybrid Clustering-Classification Neural Network in the Medical Diagnostics of the Reactive Arthritis. *IJISA* **2016**, *8*, 1–9. [[CrossRef](#)]
35. Perova, I.; Pliss, I. Deep Hybrid System of Computational Intelligence with Architecture Adaptation for Medical Fuzzy Diagnostics. *IJISA* **2017**, *9*, 12–21. [[CrossRef](#)]
36. Dhar, P.; Rahman, M.S.; Abedin, Z. Classification of Leaf Disease Using Global and Local Features. *IJITCS* **2022**, *14*, 43–57. [[CrossRef](#)]
37. Singh, A.K.; Shukla, V.P.; Biradar, S.R.; Tiwari, S. Enhanced Performance of Multi Class Classification of Anonymous Noisy Images. *IJIGSP* **2014**, *6*, 27–34. [[CrossRef](#)]
38. Heart Attack Analysis & Prediction Dataset. Available online: <https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset> (accessed on 8 May 2022).
39. Datopian Blood Transfusion Service Center. Available online: <https://datahub.io/machine-learning/blood-transfusion-service-center#data> (accessed on 6 April 2022).
40. Heart Failure Prediction. Available online: <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data> (accessed on 8 May 2022).
41. UCI Machine Learning Repository: Maternal Health Risk Data Set Data Set. Available online: <https://archive.ics.uci.edu/ml/datasets/Maternal+Health+Risk+Data+Set> (accessed on 8 May 2022).
42. UCI Machine Learning Repository: Breast Tissue Data Set. Available online: <http://archive.ics.uci.edu/ml/datasets/breast+tissue> (accessed on 8 May 2022).
43. UCI Machine Learning Repository: Contraceptive Method Choice Data Set. Available online: <https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice> (accessed on 8 May 2022).

44. Oliinyk, A.; Fedorchenko, I.; Stepanenko, A.; Rud, M.; Goncharenko, D. Implementation of Evolutionary Methods of Solving the Travelling Salesman Problem in a Robotic Warehouse. In *Data-Centric Business and Applications*; Radivilova, T., Ageyev, D., Kryvinska, N., Eds.; Lecture Notes on Data Engineering and Communications Technologies; Springer International Publishing: Cham, Switzerland, 2021; Volume 48, pp. 263–292, ISBN 978-3-030-43069-6.
45. Kumar, M.B.P.; Amaresh Savadatti, D.M. Virobot the Artificial Assistant Nurse for Health Monitoring, Telemedicine and Sterilization through the Internet. *IJWMT* **2020**, *10*, 16–26. [[CrossRef](#)]
46. Hu, Z.; Khokhlachova, Y.; Sydorenk, V.; Opirskyy, I. Method for Optimization of Information Security Systems Behavior under Conditions of Influences. *IJISA* **2017**, *9*, 46–58. [[CrossRef](#)]
47. Bykov, M.M.; Kovtun, V.V.; Smolarz, A.; Junisbekov, M.; Targeusizova, A.; Satymbekov, M. Research of Neural Network Classifier in Speaker Recognition Module for Automated System of Critical Use. In *Photonics Applications in Astronomy, Communications, Industry, and High Energy Physics Experiments*; Romaniuk, R.S., Linczuk, M., Eds.; International Society for Optics and Photonics: Wilga, Poland, 2017; p. 1044521.
48. Teslyuk, V.; Kazarian, A.; Kryvinska, N.; Tsmots, I. Optimal Artificial Neural Network Type Selection Method for Usage in Smart House Systems. *Sensors* **2021**, *21*, 47. [[CrossRef](#)]
49. Tkachenko, R. An Integral Software Solution of the SGTm Neural-Like Structures Implementation for Solving Different Data Mining Tasks. In *International Scientific Conference “Intellectual Systems of Decision Making and Problem of Computational Intelligence*; Babichev, S., Lytvynenko, V., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 696–713.