Weighted scaling approach for metabolomics data analysis

Biplab Biswas¹,² · Nishith Kumar¹ · Md. Aminul Hoque² · Md. Ashad Alam³

Received: 17 October 2022 / Revised: 14 April 2023 / Accepted: 15 April 2023 / Published online: 10 May 2023 © The Author(s) under exclusive licence to Japanese Federation of Statistical Science Associations 2023

Abstract
Systematic variation is a common issue in metabolomics data analysis. Therefore, different scaling and normalization techniques are used to preprocess the data for metabolomics data analysis. Although several scaling methods are available in the literature, however, choice of scaling, transformation and/or normalization technique influences the further statistical analysis. It is challenged to choose the appropriate scaling technique for downstream analysis to get accurate results or to make proper decision. Moreover, the existing scaling techniques are sensitive to outliers or extreme values. To fill the gap, our objective is to introduce a robust scaling approach that is not influenced by outliers as well as provides more accurate results for downstream analysis. Here, we introduced a new weighted scaling approach that is robust against outliers; however, no additional outlier detection/treatment step is needed in data preprocessing and also compared it with the conventional scaling and normalization techniques through artificial and real metabolomics datasets. We evaluated the performance of the proposed method in comparison to the other existing conventional scaling techniques using metabolomics data analysis in both the absence and presence of different percentages of outliers. Results show that in most cases, the proposed scaling technique is a better performer than the traditional scaling methods in both the absence and presence of outliers. The proposed method improves the further downstream metabolomics analysis. The R function of the proposed robust scaling method is available at https://github.com/nishithkumarpaul/robustScaling/blob/main/wscaling.R

Nishith Kumar
nishith@bsmrstu.edu.bd

Md. Ashad Alam
mdashad.alam@ochsner.org

¹ Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj 8100, Bangladesh
² Department of Statistics, Rajshahi University, Rajshahi 6205, Bangladesh
³ Centers for Outcomes and Health Services Research, Ochsner Health, 1514 Jefferson Highway, New Orleans, LA, USA
Keywords Metabolomics · Systematic variation · Outliers · Pre-treatment · Scaling · Classification accuracy

1 Introduction

In computational biology, among the omics studies, metabolomics is a newly developed and promising area that plays a connector role between genotype and phenotype. Recently, it has been playing a vital role in medical science, health science, biological research and biodiversity analysis. Biodiversity has three levels: ecological diversity, specific diversity, genetic diversity which includes different forms of life on earth (Redford et al. 2001). To assure the proper mechanism of ecosystem, species and genes, biodiversity plays an important role (Tiwari & Rajwanshi, 2022). Conservations of biodiversity are necessary for different ecosystem services like availability of wood and food, water and air purification, nutrients recycling, prevention of soil erosion, etc. (Tscharntke et al., 2012); however, there is huge damage of biodiversity due to human multiple anthropogenic activities (Opdam & Wascher, 2004). Recently, omics-assisted techniques have been introduced to conserve the biodiversity including species due to fail the different conventional conservation methods. Metabolomics data analysis using omics-assisted techniques plays a vital role for biodiversity conservations (Tiwari & Rajwanshi, 2022). Several studies were conducted on metabolomics data for biodiversity conservation such as identification of biochemical markers to investigate the progeny of non-domesticated tomato species (Schauer et al., 2005), evaluate individual plant phenotypes using metabolomics techniques (Scherling et al., 2010), assess the reasons to decrease the number of pollinators by metabolite profiling (Sardans et al., 2011), etc. Metabolomics data analysis is helpful for biological diversity, and policy makers determine the practical solutions to conserve biodiversity using the observed results.

Biological information extraction from high-dimensional, throughput molecular datasets is a major challenge in metabolomics. Metabolomics dataset is extracted from high-throughput technology through several critical steps in the data generating procedure. The most usual feature of these analytical high-throughput technologies based on nuclear magnetic resonance (NMR) and mass spectrometry (MS) is that they produce high-informative and complicated data of biological variables. Therefore, extracting important information is a confusing process (Goodacre et al., 2004), and different aspects of this type of data hamper their biological interpretation, e.g., in a metabolomics data set, there are 5000-fold differences in concentration for different metabolites (van den Berg et al., 2006). However, most metabolomics data analysis techniques are highly influenced by these differences. Therefore, data preprocessing methods like different scaling techniques can help to correctly extract biological information.

Due to huge systematic variation in metabolomics data, different scaling and normalization techniques are used to preprocess the data. These sources of variation include small changes in volume applied during sample preparation and sample injection and in instrument performance. Moreover, multi-view structure (Alam et al., 2021) can be found in the omics dataset. The above nature of the data can influence...
the biological interpretation of data. Classical statistical analysis fail to incorporate the effects of such type of structural variation as well as outliers in the metabolomics data. These also make impact on accuracy as well as other performance measures of different classification models and provide misleading inference in addition to wrong identification of metabolites in the analysis. However, robust statistical methods can consider these deviations and increases the reliability of estimates and associated inference. Thus, it is important to develop robust handling of metabolomics data analysis pipeline that will allow for the interpretation of large metabolomics datasets.

The typical pipeline of metabolomics data analysis has been described briefly in the literature (Brown et al., 2005; Goodacre et al., 2007; Sumner et al., 2007; Gromski et al., 2014; Wen, 2020), and it includes the following five steps (i) preprocessing: extracting characteristics/features from raw data to an appropriate form; (ii) pretreatment: shifting, scaling, standardization, normalization and transformation, etc.; (iii) analysis: statistical modeling as well as understanding and visualization of data; (iv) validation: predictive capability estimation of applied statistical models, and finally (v) interpretation: summary of data analysis. In data mining, as well as the metabolomics data analysis pipeline, data pretreatment is an important step among the above five steps (Goodacre et al., 2007). To interpret the biological data, the pretreatment technique plays an important role (Gromski et al., 2014), because downstream analyses entirely depend on data pretreatment techniques. It is one of the first steps in the data preparation process where the data are transformed to minimize variable redundancy as well as systematic variation that makes all variables more comparable stage (Bro & Smilde, 2003). This pretreatment process may dramatically change the result of the final output (e.g., classification accuracy, differentially concentrated expression (DE) metabolites identification).

There are several scaling techniques in the literature for pretreatment of the metabolomics dataset such as auto-scaling, vast scaling, range scaling, level scaling, Pareto scaling, etc. However, the choice/selection of scaling approaches influences the downstream data analysis (Gromski et al., 2014). Among the existing pretreatment techniques, it is found that auto-scaling, range scaling, and vast scaling are comparatively better than the other techniques (van den Berg et al., 2006; Gromski et al., 2014). However, all the above techniques are sensitive to outliers. Metabolomics datasets may contain outliers due to several steps of data quantification procedures (Blanchet & Smolinska, 2016; Kumar et al., 2017; Steuer et al., 2007). Outliers problem can be solved by (i) outliers detection and deletion/modification (Alam et al., 2018; Shahjaman et al., 2019) or (ii) developing robust functions for parameter estimation (Kumar et al., 2021). Therefore, we have developed a new weighted scaling technique using a weight function that is simple and effective and also robust to outliers, so no other outlier detection approaches are needed in data preprocessing. The performance of the proposed method has been evaluated and compared to other conventional methods using the classification accuracy by different classification techniques for both simulated and real datasets in the presence of different rates (1%, 3%, 5%, and 7%) of outliers. Experimentally measured data analyses showed that the proposed weighted scaling technique is better than other metabolomics data analysis methods.
2 Materials and methods

In this section, we have discussed the different scaling techniques, including our proposed method, and other classification techniques that have been used as a performance measure. Here, we also described the artificial and experimentally measured datasets.

2.1 Scaling techniques

This paper developed a weighted scaling technique for preprocessing the metabolomics dataset. To compare the performance of the proposed method, we considered five widely used conventional techniques: auto-scaling, range scaling, level scaling, Pareto scaling, and vast scaling, i.e., the six scaling techniques including the proposed one that has been used in this study are discussed in the following sub-section.

2.1.1 Weighted scaling (proposed)

In this section, we have discussed our proposed method, i.e., weighted scaling. It is an extension of auto-scaling using the robust version of location and scatters. To formulate the robust version of location and scatter, we modified a weight function proposed by Giloni et al. (2006). His proposal about the weight function is that the weight \( w_i \) is taken as inverse proportional to the distance from the clean subset. If \( X = (x_{ij}) \) is a metabolomics data then using the weight function the robust location and scatter can be defined as

\[
\bar{x}_i^{(w)} = \frac{\sum_j (w_{ij} x_{ij})}{\sum_j w_{ij}} \quad \text{and} \quad s_i^{(w)} = \sqrt{\frac{\sum_j (w_{ij} x_{ij} - \bar{x}_i^{(w)})^2}{\sum_j w_{ij}}},
\]

where \( w_{ij} = \min\left(1, \frac{z_{0.05}^2}{((x_{ij} - \text{median}(x_i))/\text{mad}(x_i))^2}\right) \) and \( z_{0.05} \) is the upper 5% critical value of the standard normal distribution. Where \( \text{mad} \) is the median absolute deviation \( \text{mad}(x_i) = \frac{1}{\frac{1}{0.6754} \text{median}_j \left( |x_{ij} - \text{median}_j(x_{ij})| \right)} \).

The proposed weighted scaling of \( X \) is \( \tilde{x}_{ij} = \begin{cases} 
\frac{x_{ij} - \bar{x}_i^{(w)}}{s_i^{(w)}}, & \text{if } w_{ij} = 1 \\
\frac{w_{ij} x_{ij} - \bar{x}_i^{(w)}}{s_i^{(w)}}, & \text{if } w_{ij} \neq 1 
\end{cases} \).

In the proposed method, for each observation, we took the weight in such a way that it lies between zero and one. If an observation is close to the median then \( w_{ij} \) is close to 1 and when it starts moving away from the median then \( w_{ij} \) goes towards zero. Since outliers are apart from the balk of the data points, therefore, if we use the proposed weight function the weights of outliers will be close to zero. The proposed weighted scaling approach has used the weighted mean and standard deviation. Therefore, the proposed weighted scaling is comparatively more robust against outliers and
the advantage of this method is that no other outlier detection approaches are needed in data preprocessing.

### 2.1.2 Range scaling

Range scaling scales the metabolomic concentrations by the variety of biological responses. If \( X = (x_{ij}) \) is metabolomics data then the range scaling of \( X \) is \( \tilde{X} = \left( \frac{x_{ij} - \bar{x}_i}{x_{i\text{max}} - x_{i\text{min}}} \right) \); where \( \bar{x}_i \) is the mean, \( x_{i\text{max}} \) is the maximum value and \( x_{i\text{min}} \) is the minimum value of the \( i \)-th row of \( X \).

### 2.1.3 Pareto scaling

If \( X = (x_{ij}) \) is a metabolomics data then the Pareto scaling of \( X \) is \( \tilde{X} = \left( \frac{x_{ij} - \bar{x}_i}{\sqrt{s}_i} \right) \); where \( \bar{x}_i \) and \( s_i \) are the mean and standard deviation of the \( i \)-th row of \( X \), respectively.

### 2.1.4 Vast scaling

Vast scaling is an extension of auto-scaling. In this process, the auto-scaling function is multiplied by a scaling factor divided by the standard deviation (Keun et al., 2003). If \( X = (x_{ij}) \) is a metabolomics data then the vast scaling of \( X \) is \( \tilde{X} = \left( \frac{x_{ij} - \bar{x}_i}{s_i} \right) \left( \frac{x_i}{s_i} \right) \); where \( \bar{x}_i \) and \( s_i \) are the mean and standard deviation of the \( i \)-th row of \( X \), respectively.

### 2.1.5 Level scaling

Level scaling transforms metabolic concentration relative to the average metabolic concentration by scaling according to the average metabolic concentration (van den Berg et al., 2006). If \( X = (x_{ij}) \) is a metabolomics data, then the level scaling of \( X \) is \( \tilde{X} = \left( \frac{x_{ij} - \bar{x}_i}{\bar{x}_i} \right) \); where \( \bar{x}_i \) is the mean of the \( i \)-th row of \( X \).

### 2.1.6 Auto-scaling

Auto-scaling is one of the simplest scaling methods that adjust metabolic variances (Kohl et al., 2012). In this technique, for each column, its row mean is subtracted from the whole dataset (known as centering) and also divided by the standard deviation of the corresponding row (van den Berg et al., 2006). If \( X = (x_{ij}) \) is a metabolomics data (illustrated in supplementary materials Fig. S1) then the auto-scaling of \( X \) is \( \tilde{X} = \left( \frac{x_{ij} - \bar{x}_i}{s_i} \right) \); where \( \bar{x}_i \) and \( s_i \) are the mean and standard deviation of the \( i \)-th row of \( X \), respectively.

### 2.2 Dataset description

In this article, we use two experimentally measured metabolomics datasets and an artificial dataset to investigate the performance of the proposed scaling technique compared to the prevailing five scaling methods.
2.2.1 Artificial data

In this study, we use an artificial dataset that has the cases group and control group like the experimentally measured metabolomics dataset. We consider that 106 samples come from patients with the disease and another 91 samples were taken from disease-free individuals. We designated total 236 metabolites, among these, 118 metabolites are differentially concentrated and the other 118 metabolites are non-differentially concentrated. Therefore, the dataset has 197 subjects (106 cases and 91 controls) and 236 metabolites. Data were generated using negative binomial distribution $NB(r, p)$, where $r$ is the number of successes and $p$ is the probability of success. We set $(r, p)$ as $(100, 0.60)$ and $(75, 0.55)$ for 118 differentially concentrated metabolites of 106 patients and 91 disease-free individuals, respectively, and as $(50, 0.40)$ for 118 non-differentially concentrated metabolites of all participants. In various situations, to measure the performance of the proposed scaling approach, different percentages of outliers (1%, 3%, 5%, 7%) were randomly allocated in the artificial data matrix. Outliers are generated using negative binomial distribution by changing the target for the number of successful trials and probability of success in such a way that values of the outliers are out of range of data matrix.

2.2.2 Experimentally measured data

In our study, we used two publicly accessible metabolomics datasets: breast cancer serum data and myalgic encephalomyelitis/chronic fatigue syndrome data with case and control groups. These two datasets are available at the National Institute of Health (NIH) Common Fund’s National Metabolomics Data Repository (NMDR) website. The breast cancer dataset was produced by GC–TOF-MS and processed by ChromaTOF software (v. 2.32) using the blood sample of 134 subjects. All samples were collected among the 134 subjects, 103 blood samples came from patients with breast cancer, and another 31 samples were taken from individuals without cancer. One hundred-one (101) metabolites were identified as known metabolites. Thus, the dataset contains 134 subjects (103 cancers and 31 controls) and 101 metabolites. This dataset was produced by the cancer center, the University of Hawaii, and the study ID was ST000356.

Myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS) is another dataset that was produced by Columbia University Center for Infection and Immunity under a case–control study on plasma metabolomics analysis in Myalgic encephalomyelitis/chronic fatigue syndrome with study ID: ST002003. Blood samples (plasma) were collected among 106 patients with ME/CFS and 91 samples were taken from the individuals without the disease. The dataset contains 197 subjects (106 MS/CFS and 91 controls) and 237 metabolites. To evaluate the performance of the proposed scaling approach compared to the existing scaling approaches in various scenarios, we replaced the original values of both the datasets by mixing 1%, 3%, 5%, and 7% outliers with the distribution $N(\mu, \sigma^2)$, where $\mu$ is the location parameter that is greater than the highest value of the dataset and $\sigma$ is the scale parameter which is same as the observed standard deviation of experimental data. We mixed up the outliers in such a way that 1% outliers were added to the datasets and added extra 2% outliers plus the
replaced outliers in the case of 3% outliers. This process is continued by adding 5% and 7% outliers, respectively. Moreover, we have also used a biodiversity-metabolomics data related to the movement behavior of *Salmo trutta*. Metabolomics analysis was used to observe the movement behavior (sedentary or mobile) that depends on the possible changes of metabolites of *S. trutta*. We have taken the dataset from Oromi et al., (2017) but the dataset contains lots of missing values. As data preprocessing, at first, we have deleted the metabolites which have more than 50% missing values, and then imputed the missing values using the random forest imputation technique. 1145 metabolites and 28 samples were in the dataset. There are two classes: Sedentary and Mobile.

### 2.3 Classification techniques and performance measures

To measure the performance of the proposed scaling technique compared to the other existing scaling techniques, we calculated the classification accuracy, area under receiver operating characteristics (ROC) curve (ranges between 0 and 1), F1 score, and Matthews correlation coefficient (MCC) through different classifiers like support vector machine (SVM), naïve Bayes (NB), *k*-nearest neighbor (*k*-NN) and partial least squares-discriminant analysis (PLS-DA) for the different scaled datasets (six scaling techniques have been performed for scaled dataset). MCC is mainly used for model evaluation and it measures the difference between actual values and predicted values. MCC ranges between 0 and 1. MCC = 1 indicates best agreement between actual and predicted values and 0 indicates no agreements. A short description of the calculation procedure of classification techniques is given in method section of supplementary materials S1.1–S1.4. The analyzing procedure is summarized in Fig. 1. All the data

![Fig. 1 Flowchart of the analysis procedure](image-url)
analyses performed in this paper were conducted within R (3.4.1), an open-access software environment (R Core Team 2013).

3 Results and discussion

The performance of the proposed method was compared with those of five existing scaling techniques using a generated artificial dataset, two experimentally measured datasets and the modified experimental datasets. The datasets were modified by artificially incorporating 1%, 3%, 5%, and 7% outliers. The scaling technique that produces higher accuracy (%), the area under receiver operating characteristics curve (AUC), F1 score, and Matthews correlation coefficient through four well-known classification techniques (k-NN, NB, SVM and PLS-DA) is considered a better scaling technique.

3.1 Performance evaluation for artificial data

To evaluate the performance of the proposed method, we measured the average classification accuracy rates (%) across a fivefold cross-validation with 100 iterations based on different scaling approaches for artificial data and artificial data with 1%, 3%, 5%, and 7% outliers (Case vs. Control) using k-nearest neighbor, naive Bayes, support vector machine and PLS-DA classifier. We also measured the average AUC value, average F1 score, and average MCC for all the classifiers with artificial data and modified artificial data in Table 1. From Table 1, it is observed that the proposed method gave higher values for all the above performance measures compared to other existing scaling techniques for the k-NN classifier with the scenario of a clean dataset and contamination of 1%, 3%, 5%, and 7% outliers. We also noticed that the values of performance measures decrease gradually with the increased percentage of outliers. The results of the performance indices for the artificial dataset and modified dataset for NB, SVM and PLS-DA are shown in Supplementary file 1: Table S1. The scaling approach that produces the higher accuracy rate (%), F1 score, AUC, and MCC values for a classifier is considered as a better approach. Our proposed approach gave higher values for all the performance indices that are presented in Supplementary file 1: Table S1. Similar results are also shown for the SVM and PLS-DA classifier in Table S2 and Table S3, respectively, in Supplementary file 1.

The ROC curve and performance indices (average of 100 values) across different scaling approaches with a clean artificial dataset and different percentages of outliers for the k-NN classifier are displayed in Fig. 2 and Fig. S2, respectively. In Fig. 2, our proposed scaling approach gave higher average sensitivity with respect to 1-specificity compared to the other existing scaling approaches in the absence of outliers and 1%, 3%, 5% and 7% outliers.

From Fig. S2, we also noticed that performance indices for our proposed approach are not affected by the presence of outliers; however, the performance indices values are involved in the presence of outliers for all other existing approaches. ROC curves for the NB, SVM and PLS-DA classifier are also given in Supplementary file 1 as Figs. S3, S4, and S5, respectively. Figures S3, S4, and S5 also show that the values of AUC
Table 1 Performance evaluation of different scaling approaches using k-nearest neighbor across fivefold cross-validation (100 iterations) with artificial data (Case vs. Control)

| Performance measures | Scaling techniques | Clean 1% outliers | 3% outliers | 5% outliers | 7% outliers |
|----------------------|-------------------|------------------|------------|------------|------------|
| Accuracy (%)         | Auto              | 90.74            | 86.79      | 84.09      | 82.92      | 77.37      |
|                      | Range             | 89.70            | 85.71      | 83.09      | 81.87      | 76.31      |
|                      | Level             | 89.32            | 85.92      | 83.07      | 82.38      | 76.34      |
|                      | Pareto            | 88.70            | 85.65      | 83.01      | 81.88      | 76.83      |
|                      | Vast              | 89.23            | 87.59      | 84.75      | 83.87      | 77.50      |
|                      | Proposed          | **91.99**        | **93.17**  | **92.54**  | **90.73**  | **91.21**  |
| F1 Score             | Auto              | 0.920            | 0.890      | 0.871      | 0.862      | 0.825      |
|                      | Range             | 0.910            | 0.881      | 0.862      | 0.854      | 0.818      |
|                      | Level             | 0.909            | 0.884      | 0.864      | 0.859      | 0.819      |
|                      | Pareto            | 0.904            | 0.882      | 0.863      | 0.855      | 0.822      |
|                      | Vast              | 0.908            | 0.895      | 0.875      | 0.868      | 0.825      |
|                      | Proposed          | **0.931**        | **0.940**  | **0.935**  | **0.920**  | **0.921**  |
| AUC                  | Auto              | 0.903            | 0.852      | 0.803      | 0.771      | 0.731      |
|                      | Range             | 0.892            | 0.849      | 0.859      | 0.773      | 0.710      |
|                      | Level             | 0.885            | 0.856      | 0.807      | 0.794      | 0.752      |
|                      | Pareto            | 0.880            | 0.855      | 0.785      | 0.769      | 0.745      |
|                      | Vast              | 0.879            | 0.850      | 0.840      | 0.805      | 0.721      |
|                      | Proposed          | **0.916**        | **0.923**  | **0.886**  | **0.914**  | **0.895**  |
| MCC                  | Auto              | 0.826            | 0.754      | 0.707      | 0.687      | 0.589      |
|                      | Range             | 0.798            | 0.731      | 0.686      | 0.664      | 0.569      |
|                      | Level             | 0.801            | 0.739      | 0.691      | 0.679      | 0.573      |
|                      | Pareto            | 0.789            | 0.733      | 0.689      | 0.668      | 0.579      |
|                      | Vast              | 0.794            | 0.766      | 0.717      | 0.701      | 0.589      |
|                      | Proposed          | **0.848**        | **0.868**  | **0.856**  | **0.824**  | **0.826**  |

Bold values represent the best statistically results

for our proposed approach are higher compared to the other existing conventional scaling technique in both the absence and presence of outliers.

Performance indices of different scaling approaches for the NB, SVM, and PLS-DA classifier are also given in Figs. S6, S7, and S8, respectively, in supplementary file 1. From Figs. S6, S7 and S8, it is seen that the different performance indices (Accuracy rate (%), Area under ROC curve, F1 score, Matthews correlation coefficient) of our proposed scaling method produce higher value compared to the existing scaling techniques for both the clean dataset and presence of different rates of outliers (1%, 3%, 5% and 7%). Finally, we can conclude that the proposed scaling techniques perform better than the existing five scaling techniques for artificial datasets in both presence and absence of outliers.
3.2 Performance evaluation for experimental data

Here, two experimental metabolomics datasets: breast cancer and chronic fatigue syndrome dataset were used to measure the performance of the proposed scaling approach. Using four well-known classification techniques k-nearest neighbor, Naive Bayes, support vector machine and partial least squares-discriminant analysis, we computed the different performance indices for the proposed approach and other scaling approaches using breast cancer and chronic fatigue syndrome data. We also measured the performance of the proposed scaling technique by artificially incorporating 1%, 3%, 5%, and 7% outliers in the breast cancer and chronic fatigue dataset randomly. We also measured the performance of the proposed method by identifying the differential metabolites and functional analysis using the experimental datasets.

To measure the performance of the proposed scaling techniques, we first applied the different scaling techniques in the original breast cancer and chronic fatigue dataset as well as contaminated (artificially incorporating 1%, 3%, 5%, and 7% outliers) breast...
cancer and regular fatigue datasets. Second, we classified the subjects (case vs. control) through four well-known classification techniques: k-nearest neighbor, Naive Bayes, support vector machine and partial least squares-discriminant analysis using fivefold cross-validation and also calculated the ROC curve, Accuracy, F1 score, AUC value and MCC for each dataset and each condition. We repeated it 100 times and calculated the average of each performance index. In the main manuscript, we only included the k-NN-related results. The rest of the results we put in the supplementary materials.

Performance of different scaling approaches using k-nearest neighbor across fivefold cross-validation for breast cancer dataset has been shown in Table S4, Figs. 3, and 4.

From Table S4, it is seen that in most cases, our proposed method produced higher accuracy (%), F1 Score, AUC and MCC compared to the other methods in the absence and presence of outliers. In addition, Fig. 3 exhibited that our proposed method gave a higher average TPR with respect to average FPR in comparison with the existing methods in both the absence and presence of outliers. From Fig. 4, we can see that

![Performance evaluation using ROC curve for k-nearest neighbor with different scaling methods based on breast cancer data](image)

Fig. 3 Performance evaluation using ROC curve for k-nearest neighbor with different scaling methods based on breast cancer data: (a) absence of outliers, (b) 1% outliers, (c) 3% outliers, (d) 5% outliers
Fig. 4 Comparison of different scaling approaches using performance measures for k-nearest neighbor based on breast cancer data a accuracy (%), b AUC, c F1 score, d MCC

our proposed method showed more consistent results when the outlier increases. In the breast cancer dataset, we also got similar results for Naive Bayes, support vector machine, and partial least squares-discriminant analysis that were given in supplementary materials (Tables S5–S7 and Figs. S9–S14). Therefore, for the breast cancer dataset, our proposed scaling techniques perform better than the existing five scaling techniques in both presence and absence of outliers.

On the other hand, Table S8, Figs. 5 and 6 showed the performance of different scaling approaches using k-NN across fivefold cross-validation for the chronic fatigue syndrome dataset. The values of different performance indices in Table S8 indicate that our proposed method delivers higher Accuracy (%), F1 score, AUC and MCC compared to the other methods in the absence and presence of outliers for most of the scenarios.

Moreover, Fig. 5 shows that our proposed method gave a higher average TPR with respect to average FPR in comparison with the existing methods in both the absence
Fig. 5 Performance evaluation using AUC for k-nearest neighbor with different scaling methods based on breast cancer data. 

(a) absence of outliers, 
(b) 1% outliers, 
(c) 3% outliers, 
(d) 5% outliers and presence of outliers. From Fig. 6, it is seen that our proposed method gave more consistent results than others for all the indices in the increases of outliers. In applying the Naive Bayes, support vector machine, and partial least squares-discriminant analysis on the chronic fatigue syndrome dataset, we also got similar results that were given in the supplementary materials (Tables S9–S11 and Figs. S15–S20). Consequently, it is observed that our proposed scaling techniques perform better than the existing five scaling techniques in both the absence and presence of outliers for the chronic fatigue syndrome dataset.

Moreover, we have also applied our methodology to the biodiversity-metabolomics data and results are included in the supplementary file (Table S13). These results are approximately similar in the perspectives of accuracy (%), F1 score, area under ROC curve and MCC as breast cancer and chronic fatigue syndrome metabolomics datasets. Our proposed scaling technique is better performer than the other five conventional scaling techniques.
3.3 Performance evaluation based on metabolite identification

We have identified the differentially concentrated metabolite for the breast cancer and chronic fatigue syndrome data to evaluate the performance of the proposed scaling approach over the existing scaling approaches. First, we identified the DE metabolites without preprocessing the datasets using $t$-test and fold-change techniques. Second, preprocess the datasets using existing scaling methods and our proposed method. Finally, we identified the DE metabolites from the processed data. About 34 metabolites were identified as DE metabolites through $t$-test and fold-change approach from the original dataset, i.e., without applying preprocessing methods. Similarly, we identified DE metabolites through $t$-test and fold-change approach from the five processed datasets, i.e., after applying the five existing scaling approaches based on the breast cancer dataset. After applying the proposed scaling approach breast cancer dataset, we identified one important metabolite “Cyclohexanone” as DE; however, it was not identified as a DE metabolite from the other processed datasets or clean dataset.
literature review results about “Cyclohexanone” are shown in Table S12. From Table S12, it is clear that “Cyclohexanone” is obviously liable to different types of cancers including breast cancer and lung cancer.

We also identified the DE metabolites from the chronic fatigue syndrome dataset as the same process discussed above. After applying the existing scaling approaches on the datasets, we identified 16 metabolites as DE using t-test and fold-change techniques; however, our proposed approach identified 20 metabolites as DE. Among these, Diazepam, Telmisartan, Lamotrigine, Omeprazole sulfone, Quetiapinesuloxide, R-(−)-O-Desmethylvenlafaxine, Ranitidine, Sulfamethoxazole, trans-3′-Hydroxycotlnlne, Tri-2-ethylhexyl trimellitate, Albendazole and Scopoletin were not detected by the existing scaling approaches and these metabolites are directly or indirectly associated with fatigue and other diseases that are patterns of evidence are added from the literature review in Table S12.

4 Conclusion

We have shown that the scaling technique plays an important role in classification accuracy, the area under ROC curve (AUC), F1 score, and Matthews correlation coefficients in GC–MS metabolomics data analysis. We have also shown that outlying observations strongly influence the performance of existing scaling techniques. Therefore, we have proposed a robust scaling technique using a weight function in this paper. We investigated the performance of the proposed method in a comparison of the traditional five methods (Auto Scaling, Vast Scaling, Level Scaling, Pareto Scaling and Range Scaling) using an artificially generated metabolomics dataset and experimentally measured breast cancer and chronic fatigue syndrome datasets. Based on our computational findings, we have concluded that the proposed scaling technique is a better performer than the traditional scaling techniques in both the absence and presence of different rates (1%, 3%, 5%, and 7%) of outliers. Therefore, we recommend using our proposed robust scaling technique instead of existing methods to scale the GC–MS metabolomics data for further univariate, multivariate, and exploratory metabolomics data analysis.

Supplementary Information  The online version contains supplementary material available at https://doi.org/10.1007/s42081-023-00205-2.

Acknowledgements  We thanks and acknowledge the Ministry of Science and Technology, Bangladesh for the National Science and Technology (NST) fellowship.

Author contributions  BB analyzed the data, drafted the manuscript, and executed the statistical analysis. NK and MAA worked to develop the weighted scaling approach for metabolomics data analysis. NK, MAA, and MAH coordinated and supervised the project. All authors carefully read and finally approved the manuscript.

Data availability  Two datasets are available at the National Institute of Health (NIH) Common Fund’s National Metabolomics Data Repository (NMDR) website (https://www.metabolomicsworkbench.org/data/index.php). The breast cancer dataset was produced by GC–TOF-MS and processed by ChromaTOF software (v. 2.32) using the blood sample of 134 subjects.
Declarations

Conflict of interest The authors declare that they have no known competing financial interests.

References

Alam, M. A., Calhoun, V. D., & Wang, Y. P. (2018). Identifying outliers using multiple kernel canonical correlation analysis with application to imaging genetics. *Computational Statistics & Data Analysis, 125*, 70–85. https://doi.org/10.1016/j.csda.2018.03.013

Alam, M. A., Qiu, C., Shen, H., Wang, Y. P., & Deng, H. W. (2021). A generalized kernel machine approach to identify higher-order composite effects in multi-view datasets, with application to adolescent brain development and osteoporosis. *Journal of Biomedical Informatics, 120*, 103854. https://doi.org/10.1016/j.jbi.2021.103854

Benbadis, S., Klein, P., Schiemann, J., Diaz, A., Elmoufi, S., & Whitesides, J. (2018). Efficacy, safety, and tolerability of brivaracetam with concomitant lamotrigine or concomitant topiramate in pooled phase III randomized, double-blind trials: A post-hoc analysis. *Epilepsy & Behavior, 80*, 129–134. https://doi.org/10.1016/j.yebeh.2017.12.024

Bhajni, E., Sehgal, V. K., Kumar, A., & Sehgal, A. (2020). The comparative study of azilsartan with telmisartan in terms of efficacy safety and cost-effectiveness in hypertension. *International Journal of Medical and Dental Sciences*. https://doi.org/10.18311/ijmds/2020/24723

Blanchet, L., & Smolinska, A. (2016). Data fusion in metabolomics and proteomics for biomarker discovery. *Statistical Analysis in Proteomics, 1362*, 209–223. https://doi.org/10.1007/978-1-4939-3106-4_14

Brown, M., Dunn, W. B., Ellis, D. I., Goodacre, R., Handl, J., Knowles, J. D., O’Hagan, S., Spasić, I., & Kell, D. B. (2005). A metabololome pipeline: From concept to data to knowledge. *Metabolomics, 1*(1), 39–51. https://doi.org/10.1007/s11306-005-1106-4

Dialiwal, J. S., Rosani, A., & Saadabadi, A. (2021). Diazepam. In *StatPearls*. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK537022/

Fordellone, M., Bellincontro, A. & Mencarelli, F. (2018). Partial least squares discriminant analysis: A dimensionality reduction method to classify hyperspectral data. *arXiv preprint arXiv:1806.09347*. https://doi.org/10.48550/arXiv.1806.09347

Giloni, A., Simonoff, J. S., & Sengupta, B. (2006). Robust weighted LAD regression. *Computational Statistics and Data Analysis, 50*, 3124–3140. https://doi.org/10.1016/j.csda.2005.06.005

Goodacre, R., Broadhurst, D., Smilde, A. K., Kristal, B. S., Baker, J. D., Berger, R., Bessant, C., Connor, S., Capuani, G., Craig, A., & Ebbels, T. (2007). Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics, 3*(3), 231–241. https://doi.org/10.1007/s11306-007-0081-3

Goodacre, R., Vaidyanathan, S., Dunn, W. B., Harrigan, G. G., & Kell, D. B. (2004). Metabolomics by numbers: Acquiring and understanding global metabolite data. *Trends in Biotechnology, 22*, 245–252. https://doi.org/10.1016/j.tibtech.2004.03.007

Gromski, P. S., Xu, Y., Kotze, H. L., Correa, E., Ellis, D. I., Armitage, E. G., Turner, M. L., & Goodacre, R. (2014). Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites, 4*(2), 433–452. https://doi.org/10.3390/metabo4020433

Guo, L., Wang, C., Chi, C., Wang, X., Liu, S., Zhao, W., Ke, C., Xu, G., & Li, E. (2015). Exhaled breath volatile biomarker analysis for thyroid cancer. *Translational Research, 166*(2), 188–195. https://doi.org/10.1016/j.trsl.2015.01.005. 10.1016/j.yebeh.2017.12.024.

Janssens, E., van Meerbeek, J. P., & Lamote, K. (2020). Volatile organic compounds in human matrices as lung cancer biomarkers: A systematic review. *Critical Reviews in Oncology/hematology, 153*, 103037. https://doi.org/10.1016/j.critrevonc.2020.103037
Keun, H. C., Ebbels, T. M., Antti, H., Bollard, M. E., Beckonert, O., Holmes, E., Lindon, J. C., & Nicholson, J. K. (2003). Improved analysis of multivariate data by variable stability scaling: Application to NMR-based metabolic profiling. *Analytica Chimica Acta*, 490, 265–276. https://doi.org/10.1016/S0003-2670(03)00094-1

Kim, C. H., Kang, S. I., & Shin, D. (2021). Pharmacokinetic interaction between Telmisartan and rosuvastatin/ezetimibe after multiple oral administration in healthy subjects. *Advances in Therapy*, 38(2), 1094–1105. https://doi.org/10.1007/s12325-020-01592-8

Kohl, S. M., Klein, M. S., Hochrein, J., Oefner, P. J., Spang, R., & Gronwald, W. (2012). State-of-the-art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics*, 8(1), 146–160. https://doi.org/10.1007/s11306-011-0350-z

Kumar, N., Hoque, M. A., Shahjaman, M., Islam, S. M. S., & Mollah, M. N. H. (2017). Metabolomic biomarker identification in presence of outliers and missing values. *BioMed Research International*, 2017, 1–11. https://doi.org/10.1155/2017/2437608

Kumar, N., Hoque, M., & Sugimoto, M. (2021). Kernel weighted least square approach for imputing missing values of metabolomics data. *Scientific Reports*, 11(1), 1–12. https://doi.org/10.1038/s41598-021-90654-0

Leung, E., Newcastle, G. W., Joseph, W. R., Rosengren, R. J., Larsen, L., & Baguley, B. C. (2012). Identification of cyclohexanone derivatives that act as catalytic inhibitors of topoisomerase I: Effects on tamoxifen-resistant MCF-7 cancer cells. *Investigational New Drugs*, 30(6), 2103–2112. https://doi.org/10.1007/s10637-011-9768-4

Li, B., Tang, J., Yang, Q., Cui, X., Li, S., Chen, S., Cao, Q., Xue, W., Chen, N., & Zhu, F. (2016a). Performance evaluation and online realization of data-driven normalization methods used in lc/ms based untargeted metabolomics analysis. *Scientific Reports*. https://doi.org/10.1038/srep38881

Li, Z., Yang, C., Liu, K., Hu, F., & Jin, B. (2016b). Automatic scaling hadoop in the cloud for efficient process of big geospatial data. *ISPRS International Journal of Geo-Information*, 5(10), 173. https://doi.org/10.3390/ijgi5100173

Lima, A. R., Araújo, A. M., Pinto, J., Jerónimo, C., Henrique, R., Bastos, M. D. L., Carvalho, M., & Guedes de Pinho, P. (2018). Discrimination between the human prostate normal and cancer cell exometabolome by GC-MS. *Scientific Reports*, 8(1), 1–12. https://doi.org/10.1038/s41598-018-23847-9

Liu, H., Wang, H., Li, C., Wang, L., Pan, Z., & Wang, L. (2014). Investigation of volatile organic metabolites in lung cancer pleural effusions by solid-phase microextraction and gas chromatography/mass spectrometry. *Journal of Chromatography B*, 945, 53–59. https://doi.org/10.1016/j.jchromb.2013.11.038

Mochalski, P., King, J., Haas, M., Unterkofler, K., Amann, A., & Mayer, G. (2014). Blood and breath profiles of volatile organic compounds in patients with end-stage renal disease. *Bmc Nephrology*, 15(1), 1–14. https://doi.org/10.1186/1471-2369-15-43

Negro, A., De Marco, L., Cesario, V., Santi, R., Boni, M. C., & Zanelli, M. (2017). A case of moderate sprue-like enteropathy associated with telmisartan. *Journal of Clinical Medicine Research*, 9(12), 1022. https://doi.org/10.14740/jocmr3047w

Nie, J. M., & Li, H. F. (2018). Therapeutic effects of Salvia miltiorrhiza injection combined with telmisartan in patients with diabetic nephropathy by influencing collagen IV and fibronectin: A case-control study. *Experimental and Therapeutic Medicine*, 16(4), 3405–3412. https://doi.org/10.3892/etm.2018.6654

Opdam, P., & Wascher, D. (2004). Climate change meets habitat fragmentation: Linking landscape and biogeographical scale levels in research and conservation. *Biological Conservation*, 117(3), 285–297. https://doi.org/10.1016/j.biocon.2003.12.008

Oromi, N., Jove, M., Pascual-Pons, M., Royo, J. L., Rocaspana, R., Aparicio, E., Pamplona, R., Palau, A., Sanuy, D., Fibrna, J., & Portero-Otin, M. (2017). Differential metabolic profiles associated to movement behaviour of stream-resident brown trout (Salmo trutta). *PLoS ONE*, 12(7), e0181697. https://doi.org/10.1371/journal.pone.0181697

Redford, K. H., & Richter, B. D. (1999). Conservation of biodiversity in a world of use. *Conservation Biology*, 13(6), 1246–1256. https://doi.org/10.1046/j.1523-1739.1999.97463.x

Sardans, J., Penuelas, J., & Rivas-Ubach, A. (2011). Ecological metabolomics: Overview of current developments and future challenges. *Chemoecology*, 21(4), 191–225. https://doi.org/10.1007/s00049-011-0083-5

Schauer, N., Zamir, D., & Fernie, A. R. (2005). Metabolic profiling of leaves and fruit of wild species tomato: A survey of the Solanum lycopersicum complex. *Journal of Experimental Botany*, 56(410), 297–307. https://doi.org/10.1093/jxb/eri057
Scherling, C., Roscher, C., Giavalisco, P., Schulze, E. D., & Weckwerth, W. (2010). Metabolomics unravel contrasting effects of biodiversity on the performance of individual plant species. PLoS ONE, 5(9), e12569. https://doi.org/10.1371/journal.pone.0012569

Shahjaman, M., Kumar, N., & Mollah, M. N. (2019). Performance improvement of gene selection methods using outlier modification rule. Current Bioinformatics, 14(6), 491–503. https://doi.org/10.2174/1574893614666181126110008

Silva, C. L., Perestrelo, R., Silva, P., Tomás, H., & Câmara, J. S. (2017). Volatile metabolomic signature of human breast cancer cell lines. Scientific Reports, 7(1), 1–8. https://doi.org/10.1038/srep43969

Skryabin, V. Y., Zastrozhin, M., Torrado, M., Grishina, E., Ryzhikova, K., Shipitsyn, V., Galaktionova, T., Sorokin, A., Bryun, E., & Sychev, D. (2021). Effects of CYP2C19*17 genetic polymorphisms on the steady-state concentration of diazepam in patients with alcohol withdrawal syndrome. Hospital Pharmacy, 56(5), 592–596. https://doi.org/10.1177/0018578720931756

Smeralda, C. L., Gigli, G. L., Janes, F., & Valente, M. (2020). May lamotrigine be an alternative to topiramate in the prevention of migraine with aura? Results of a retrospective study. BMJ Neurology Open. https://doi.org/10.1136/bmjno-2020-000059

Steuer, R., Morgenthal, K., Weckwerth, W., & Selbig, J. (2007). A gentle guide to the analysis of metabolomic data. In W. Weckwerth (Ed.), Metabolomics. Methods in molecular biology™. (Vol. 358). Humana Press. https://doi.org/10.1007/978-1-59745-244-1_7

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W. M., Fiehn, O., Goodacre, R., Griffin, J. L., & Hankemeier, T. (2007). Proposed minimum reporting standards for chemical analysis. Metabolomics, 3(3), 211–221. https://doi.org/10.1007/s11306-007-0082-2

Tiwari, S., & Rajwanshi, R. (2022). Overview of omics-assisted techniques for biodiversity conservation. In A. Kumar, B. Choudhury, S. Dayanandan, & M. L. Khan (Eds.), Molecular genetics and genomics tools in biodiversity conservation. Springer. https://doi.org/10.1007/978-981-16-6005-4_4

Tschahrntke, T., Batáry, P., Clough, Y., Kleijn, D., Scherber, C., Thies, C., Wanger, T.C. & Westphal, C. (2012). Combining biodiversity conservation with agricultural intensification. Land use intensification. Effects on agriculture, biodiversity and ecological processes, CSIRO Publishing: Australia, 7–15.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: Improving the biological information content of metabolomics data. BMC Genomics, 7(1), 142. https://doi.org/10.1186/1471-2164-7-142

Wang, C., Sun, B., Guo, L., Wang, X., Ke, C., Liu, S., Zhao, W., Luo, S., Guo, Z., Zhang, Y., & Xu, G. (2014). Volatile organic metabolites identify patients with breast cancer, cyclomastopathy and mammary gland fibroma. Scientific Reports, 4(1), 1–6. https://doi.org/10.1038/srep05383

Wen, B. (2020). Modular metaX pipeline for processing untargeted metabolomics data. Processing Metabolomics and Proteomics Data with Open Software. https://doi.org/10.1039/9781788019880-00302

Wood, K. E., Palmer, K. L., & Krasowski, M. D. (2021). Correlation of elevated lamotrigine and levetiracetam serum/plasma levels with toxicity: A long-term retrospective review at an academic medical center. Toxicology Reports, 8, 1592–1598. https://doi.org/10.1016/j.toxrep.2021.08.005

Worden, L., Grocott, O., Tourjee, A., Chan, F., & Thibert, R. (2018). Diazepam for outpatient treatment of nonconvulsive status epilepticus in pediatric patients with Angelman syndrome. Epilepsy & Behavior, 82, 74–80. https://doi.org/10.1016/j.ybeh.2018.02.027

Zhao, X., Huang, X., Peng, W., Han, M., Zhang, X., Zhu, K., & Shao, B. (2022). Chlorine disinfection byproduct of diazepam affects nervous system function and possesses gender-related difference in zebrafish. Ecotoxicology and Environmental Safety, 238, 113568. https://doi.org/10.1016/j.ecoenv.2022.113568

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.