# Kernel Weighted Least Square Approach for Imputing Missing Values of Metabolomics Data

**Nishith Kumar** (nk.bru09@gmail.com)
Bangabandhu Sheikh Mujibur Rahman Science and Technology University

**Md. Hoque**
University of Rajshahi

**Masahiro Sugimoto**
Tokyo Medical University

**Research Article**

**Keywords:** Metabolomics, Missing data imputation, Weighted least square, Receiver operating characteristic (ROC) curve, Area under the ROC curve (AUC), Support vector machine (SVM)

**Posted Date:** January 13th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-140282/v1

# Kernel Weighted Least Square Approach for Imputing Missing Values of Metabolomics Data

Nishith Kumar[1*], Md. Aminul Hoque[2] , Masahiro Sugimoto[3]

[1]Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh

[2]Department of Statistics, University of Rajshahi, Rajshahi, Bangladesh

[3]Health Promotion and Preemptive Medicine, Research and Development Center for Minimally Invasive Therapies, Tokyo Medical University, Shinjuku, Tokyo, 160-8402, Japan

***Correspondence:** Nishith Kumar

Department of Statistics, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Gopalganj, Bangladesh, E-mail: nk.bru09@gmail.com; Tel.: +88-01925200899

**Keywords:** Metabolomics, Missing data imputation, Weighted least square, Receiver operating characteristic (ROC) curve, Area under the ROC curve (AUC), Support vector machine (SVM).

**Running Title:** Kernel Weighted Least Square for Missing Value Imputation

## Abstract

Mass spectrometry is a modern and sophisticated high-throughput analytical technique that enables large-scale metabolomics analyses. It yields a high dimensional large scale matrix (samples × metabolites) of quantified data that often contain missing cell in the data matrix as well as outliers which originate from several reasons, including technical and biological sources. Although, in the literature, several missing data imputation techniques can be found, however all the conventional existing techniques can only solve the missing value problems but not relieve the problems of outliers. Therefore, outliers in the dataset, deteriorate the accuracy of imputation. To overcome both the missing data imputation and outlier's problem, here, we developed a new kernel weight function based missing data imputation technique (proposed) that resolves both the missing values and outliers. We evaluated the performance of the proposed method and other nine conventional missing imputation techniques using both artificially generated data and experimentally measured data analysis in both absence and presence of different rates of outliers. Performance based on both artificial data and real metabolomics data indicates that our proposed kernel weight based missing data imputation technique is a better performer than some existing alternatives. For user convenience, an R package of the proposed kernel weight based missing value imputation technique has been developed which is available at https://github.com/NishithPaul/tWLSA .

## 1. Introduction

Metabolomics datasets produced by mass spectrometry (MS) technique often contain a wide number of missing cell in the data matrix that can generate from various sources including both technological and biological hazard. Generally, there are about 10% to 40% missing values in metabolomics dataset[1-3]. The following reasons can be observed for occurring missing values in metabolomics dataset: (i) metabolite concentration peak is below the analytical method's detectable threshold (ii) metabolite concentration peak is not initially present in the chromatogram (iii) due to overlapping signals separation (iv) deconvolution may give false negative during separation of overlapping signals (v) computational and /or measurement error, (vi) the concentration of metabolite is existent in the sample but is vanished during downstream processing and finally (vii) the concentration of a particular metabolite is identified in one sample, however, is not existent at significant concentration in another sample[1,3-6]. Furthermore, the above missingness can be categorized as (a) MCAR-missing completely at random (b) MAR-missing at random (c) MNAR-missing not at random. If a missing is not related with any observed variable or response is called MCAR, whereas, if a missing is linked with one or more observed variables but not to the response is said to be MAR, finally the response associated missing is said to be MNAR. In metabolomics dataset, if the concentration of a metabolite is not seen in one group of samples but is present in another group of samples, then the missing are most probably occurring for a biological reason and can be classified as MNAR. However, if the peak of metabolite concentration is smaller than the analytical method's detectable threshold, then this type of missing is the combination of biological and technological issues, therefore, this type of missing can be considered as MNAR. Finally MCAR is caused by only technological reasons, e.g., errors related with peak picking software, i.e., peak was shown; however, it was not included in the raw data.

The easiest and state forward method of dealing with missing values is filtering method, whereby, variables[7, 8] or samples[9, 10] are removed. In recent times this is rarely applicable only when the data matrix includes a greater percentage of missing data. To handle missing value problem, the alternative approach is imputation technique. Current imputation techniques for imputing missing data are half of the minimum value replacement[2,11], mean replacement[12], median replacement[12], k-nearest neighbour (kNN)[13], Bayesian principal component analysis (BPCA)[14,15], probabilistic principal component analysis (PPCA)[16], zero imputation[17], multiple imputation with expectation maximization (EM) algorithm and monte carlo markov chain (MCMC) method[18], expectation maximization principal component analysis (EM-PCA)[19], random forest (RF) imputation[20] and so on. In the literature, there are several missing imputation techniques, however, the selection of missing imputation technique play a dramatic impact on univariate and multivariate (unsupervised and supervised) data analysis and its interpretation[1,21-23]. Therefore, the appropriate handling of missing data is very important according to the structure or nature of original data for downstream analysis. The pattern of metabolomics dataset is very complicated, because metabolomics dataset contains outliers[24], non-normality and inherent correlation structure[25]. All the aforementioned techniques can only deal with the problem of missing value imputation. However, all the existing missing value imputation techniques are more or less influenced by outliers as well as cannot significantly reduced the outlier's problem simultaneously, because the conventional imputation algorithms didn't consider any outlier-robust function or any outlier identification and substitute algorithms directly. Furthermore, the existing outliers resolving techniques do not consider with missing value problems. For these reasons, here, we have developed a novel kernel weight based missing imputation (KMI) method that can overcome both the missing value imputation problems and outliers simultaneously.

To evaluate the performance of the introduced missing imputation method with the others conventional missing value imputation methods, we took into account nine well-known and currently used missing imputation methods: zero imputation, mean imputation, median imputation, half of the minimum value imputation, kNN imputation, BPCA imputation, PPCA imputation, EM-PCA imputation and random forest (RF) imputation. We measured the performances of the missing imputation methods including the proposed one technique using both artificial and real data analysis in absence and presence of different rates of outliers.

## 2. Material and Methods

In this dissertation, we developed a new missing data imputation method by minimizing two ways kernel weighted square error loss function. To compare the competence of the proposed method, we considered nine traditional missing imputation technique: zero, mean, median, half of the minimum value, kNN, BPCA, PPCA, EM-PCA and random forest (RF) imputations. If all the missing values are substituted by zero is known as zero imputation. In mean, median and half of the minimum value imputation, missing data of each metabolite are substituted by the corresponding metabolite average, median and half of the minimum value respectively. Missing data substitution using kNN, EM-PCA, and RF are found in the *"impute"*, *"missMDA"* and *"missForest"* packages respectively of R platform. Moreover, BPCA and PPCA imputation can be done using *"pcaMethods"* package in Bioconductor. The detail description of the proposed missing value imputation method by using two-way kernel weighted square error loss function is given below,

### Missing data imputation using two-way kernel weighted least square error approach (Proposed)

Let $X = \left(x_{ij}\right)$is a metabolomics data; where, $i = 1, 2, \cdots, p$ represent the metabolites and $j = 1, 2, \cdots, n$ represents the samples. Thus, in the metabolomics data *X,* different rows indicates different metabolites and the columns indicate different samples.

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix}$$

Each cell of the metabolomics data could be represented as the product of the metabolite (row) effect and sample (column) effect. Mathematically, it is written as the bilinear form,

$$x_{ij} = r_i c_j \tag{1}$$

where, $r_i$ and $c_j$ represent the $i$-th row effect (i.e., metabolite effect) and $j$-th column effect (i.e., sample effect) respectively. Since, observed metabolomics data matrix usually contain missing cell and outliers, therefore, both the missing cell and outliers in the data matrix can be estimated by considering the effect of corresponding row and column. In equation (1), $r_i$ and $c_j$ are both unknown therefore, our motive is to determine $r_i$ and $c_j$ to forecast the $ij$-th missing cell or outlying cell. To estimate $r_i$ and $c_j$, let us consider the model $x_{ij} = r_i c_j + \in_{ij}$ (2); where, $x_{ij}$ is the yield corresponding to the effect of $i$th metabolite (row) and $j$-th sample (column), $r_i$ indicate the factors of $i$th metabolite and $c_j$ indicate the factors of $j$-th sample and $\in_{ij}$ indicate the error term. From the model (2), we have to estimate $r_i$ and $c_j$ simultaneously. To estimate $r_i$ and $c_j$ we have developed the weighted least square approach using a kernel weight function $w_j = \exp\{-\frac{\lambda}{2(mad(x_j))^2}(x_{ij} - median(x_j))^2\}$ and updated $r_i$ and $c_j$ by iterative procedure, where, $mad$ stands for median absolute deviation. The specialty of the kernel weight function is that it lies between zero and one; the weight will be closed to zero if the corresponding observation is apart from its median and also closed to one if the corresponding observation is the neighbor of median. In the kernel weight function, $\lambda$ is the tuning parameter, where the value of $\lambda$ is chosen by the k-fold cross validation (Diagram 1 indicates the appropriate $\lambda$ selection procedure). If the dataset is clean (i.e., no outliers) then $\lambda$ will be zero, at that time all the weights will be 1 i.e., the technique will be classical least square approach. The steps of estimating $r_i$ and $c_j$ are given below,

**Step-1**. To initialize the $j$-th column (sample) effect ($c_j$), calculate the $j$-th column median of $X$. Column median is computed by excluding the missing values for $j = 1, 2, \cdots, n$.

**Step-2**. Using the weighted least square approach, estimate the $i$-th row effect (i.e., metabolite effect) $r_i$ by minimizing $\sum_{j=1}^{n}\left(e_{ij}\right)^2 = \sum_{j=1}^{n} w_{ij}\left(x_{ij} - r_i c_j\right)^2$, based on $i$-th row of $X$, by eliminating the missing values, $i = 1, 2, \cdots, p$.

**Step-3**. Revise the *j*-th column effect $c_j$, using the weighted least square approach by minimizing $\sum_{i=1}^{p} w_{ij}\left(x_{ij} - r_i c_j\right)^2$, based on *j*-th column of *X*, by eliminating the missing values, $j = 1, 2, \cdots, n$.

**Step 4.** Repeat Step 2 and Step 3 until it satisfies the rule $\frac{|\boldsymbol{r}_{new} - \boldsymbol{r}_{old}| + |\boldsymbol{c}_{new} - \boldsymbol{c}_{old}|}{n+p} \leq \varepsilon$; here $\varepsilon$ is a very small positive number, it depends on researcher interest. Here, we choose $\varepsilon = 0.01$.

**Step 5.**

- Compute the first fitted bilinear form as, $\hat{X}^{(1)} = \hat{\boldsymbol{r}}_1 \hat{\boldsymbol{c}}_1$ , where $\hat{\boldsymbol{r}}_1 = (\hat{r}_1, \hat{r}_2, \cdots, \hat{r}_p)^T$ and $\hat{\boldsymbol{c}}_1 = (\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_n)$ are obtained from Step 4.
- Calculate the first remainder matrix (*X$_{R1}$*) as, $X_{R1} = X - \hat{X}^{(1)} = X - \hat{\boldsymbol{r}}_1 \hat{\boldsymbol{c}}_1$ (excluding the missing cells of the data matrix)
- Using steps 1-4 on *X$_{R1}$*, compute the second fitted bilinear form as, $\hat{X}_{R1} = \hat{\boldsymbol{r}}_2 \hat{\boldsymbol{c}}_2$ and calculate the second remainder matrix (*X$_{R2}$*) as, $X_{R2} = X_{R1} - \hat{X}_{R1} = X - \hat{\boldsymbol{r}}_1 \hat{\boldsymbol{c}}_1 - \hat{\boldsymbol{r}}_2 \hat{\boldsymbol{c}}_2$ (excluding the missing cells of the data matrix)
- Similarly calculate the *r*-th remainder (*X$_{Rr}$*) as, $X_{Rr} = X_{R(r-1)} - \hat{X}_{R(r-1)} = X - \sum_{k=1}^{r} \hat{\boldsymbol{r}}_k \hat{\boldsymbol{c}}_k$ i.e., $X = X_{Rr} + \sum_{k=1}^{r} \hat{\boldsymbol{r}}_k \hat{\boldsymbol{c}}_k$ . The number of *r* is selected in such a way that the total row variations of $\sum_{k=1}^{r} \hat{\boldsymbol{r}}_k \hat{\boldsymbol{c}}_k$ can explain (1-*α*)100% variations of *X* (using the concept of singular value decomposition; the detail of *r* selection procedure is given in Appendix 1 of supplementary materials), where *α* is chosen by the researcher interest. In this case, we took *α*=0.05. Therefore, the approximation of $X$ is,

$$X \approx \hat{X}^{(r)} = \sum_{k=1}^{r} \hat{\boldsymbol{r}}_k \hat{\boldsymbol{c}}_k \qquad (3)$$

**Step 6.** Substitute the missing values and the outlying cells of $X$ by the corresponding cells of $\hat{X}^{(r)}$ that produce the reconstructed full and clean data matrix $\tilde{X}^{c}$. Here, inter quartile range (IQR) rule[26] is used to detect outliers.

The application procedure of the proposed method in metabolomics data is given below,

Metabolomics dataset may contain several groups of samples in their data structure, if a metabolomics dataset contain $k$ groups of samples, then the dataset is split according to the groups as,

$$X = \begin{bmatrix} \overbrace{\begin{matrix} x_{11} & x_{12} & \cdots & x_{1g_1} \\ x_{21} & x_{22} & \cdots & x_{2g_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pg_1} \end{matrix}}^{group\text{-}1} & \overbrace{\begin{matrix} x_{1(g_1+1)} & x_{1(g_1+2)} & \cdots & x_{1(g_1+g_2)} \\ x_{2(g_1+1)} & x_{2(g_1+2)} & \cdots & x_{2(g_1+g_2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p(g_1+1)} & x_{p(g_1+2)} & \cdots & x_{p(g_1+g_2)} \end{matrix}}^{group\text{-}2} & \overbrace{\begin{matrix} \cdots \\ \cdots \\ \ddots \\ \cdots \end{matrix}}^{\cdots} & \overbrace{\begin{matrix} x_{1(g_1+\cdots+g_{k-1}+1)} & x_{1(g_1+\cdots+g_{k-1}+2)} & \cdots & x_{1(g_1+\cdots+g_k)} \\ x_{2(g_1+\cdots+g_{k-1}+1)} & x_{2(g_1+\cdots+g_{k-1}+2)} & \cdots & x_{2(g_1+\cdots+g_k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p(g_1+\cdots+g_{k-1}+1)} & x_{p(g_1+\cdots+g_{k-1}+2)} & \cdots & x_{p(g_1+\cdots+g_k)} \end{matrix}}^{group\text{-}k} \end{bmatrix}$$

; where, $g_1$ is the column number (subjects) of group-1, $g_2$ is the column number (subjects) of group-2 and so on; and also $g_1 + g_2 + \cdots + g_k = n$.

Therefore, check whether the metabolomics data matrix $X$ contains multiple groups or not in samples. If $X$ contains multiple groups, then partition the matrix $X$ as, $X = (X_1 \quad X_2 \quad \cdots \quad X_k)$ according to $k$ groups of samples,

where, $X_1 = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1g_1} \\ x_{21} & x_{22} & \cdots & x_{2g_1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pg_1} \end{pmatrix}$, $X_2 = \begin{pmatrix} x_{1(g_1+1)} & x_{1(g_1+2)} & \cdots & x_{1(g_1+g_2)} \\ x_{2(g_1+1)} & x_{2(g_1+2)} & \cdots & x_{2(g_1+g_2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p(g_1+1)} & x_{p(g_1+2)} & \cdots & x_{p(g_1+g_2)} \end{pmatrix}$ and

$$X_k = \begin{pmatrix} x_{1(g_1+\cdots+g_{k-1}+1)} & x_{1(g_1+\cdots+g_{k-1}+2)} & \cdots & x_{1(g_1+\cdots+g_k)} \\ x_{2(g_1+\cdots+g_{k-1}+1)} & x_{2(g_1+\cdots+g_{k-1}+2)} & \cdots & x_{2(g_1+\cdots+g_k)} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p(g_1+\cdots+g_{k-1}+1)} & x_{p(g_1+\cdots+g_{k-1}+2)} & \cdots & x_{p(g_1+\cdots+g_k)} \end{pmatrix};$$

$$\text{otherwise, } X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{pmatrix}$$

If $X$ contains $k$ groups, then apply Steps 1-6 for each partitioned data matrix and compute $\tilde{X}_1, \ \tilde{X}_2, \ \cdots \ \tilde{X}_k$; thus, the reconstructed full and clean data matrix $\tilde{X} = (\tilde{X}_1 \ \ \tilde{X}_2 \ \ \cdots \ \ \tilde{X}_k)$. Otherwise, apply Steps 1-6 for the data matrix $X$ and compute the reconstructed full and clean data matrix $\tilde{X}$.

User can install the package in R platform using the following R code

```
library(devtools)
install_github("NishithPaul/tWLSA")
library(tWLSA)
```

**Artificially Generated Metabolomics Data**

To simulate metabolomics datasets, we used the following additive linear model as,

$$x_{ijk} = \mu_i + g_{ij} + \in_{ijk} \qquad (4),$$

where, $x_{ijk}$ is the concentration of the $i^{\text{th}}$ metabolite, $j^{\text{th}}$ group and $k^{\text{th}}$ sample; the average concentration for the *i*-th metabolite is $\mu_i$; $g_{ij}$ represents the $j^{\text{th}}$ group effect of $i^{\text{th}}$ metabolite and the random error term of the *i*-th metabolite, *j*-th group and *k*-th sample is $\in_{ijk}$. To generate the data, we considered $\mu_i \sim uniform(5,10)$ and $\in_{ijk} \sim N(0,1)$. To measure the efficiency of the proposed technique, we made three types of metabolomics datasets, (i) without class level in the samples (ii) two class levels (two groups) in the samples (iii) three class levels (three groups) in the samples. In case of two class and three class level based datasets, we also generated two types of metabolites (a) equal concentration (EE) metabolites (b) differential concentrations (DE) metabolites. DE metabolites were considered into two groups: up-concentrated and down concentrated metabolites. For up-concentrated metabolites, we took $g_{ij} \sim N(0,1)$ for healthy group and $g_{ij} \sim N(2,1)$ for disease group. Similarly, for down concentrated metabolites, we took $g_{ij} \sim N(2,1)$ for healthy group and $g_{ij} : N(0,1)$ for disease group. In EE metabolites, $g_{ij} : N(0,1)$ for both groups. We generated 200 metabolites and 90 samples for each dataset. In two class and three class datasets we considered 80 metabolites as DE and 120 metabolites as EE. We generated 100 datasets for each type. We also incorporated various rates (5%, 10%, 15% and 20%) of missing cells in the data matrix. Among the total missing, 60% MAR and 40% for lower values. To investigate the efficiency of our proposed technique in presence of outliers, we also included the various rates (3%, 5%, 7% and 10% ) of outliers in the artificial datasets. In the *i*-th metabolite, we provided $N(5*\mu_i ,\sigma_i^2 )$ *as* outliers; where $\mu_i$ and $\sigma_i^2$ were the *i*-th metabolite's mean and variance; these outliers were distributed randomly in the dataset, thus, outliers may occur anywhere in the dataset.

**Real Metabolomics Data**

To measure the performance of our proposed missing imputation method firstly we consider publicly available two fully defined real metabolomics data matrix from R language platform, one is on Human Cachexia dataset (available in R-specmine library) and treated dataset (available in R-metabolomics library). Since, these two data matrices didn't contain any missing values, therefore, to investigate the efficiency of the proposed technique compared to the other techniques; we randomly incorporated different rates (5%, 10%, 15% and 20%) of missing values and also computed the mean square error (MSE) between the reconstructed data and original data. We also considered two datasets-Hepatocellular Carcinoma (HCC) with 26.52% missing values/cells[27] and MDA-MB-231 breast cancer dataset with 15.81% missing values[28] for evaluating the performance of the proposed missing value imputation method. HCC and MDA-MB-231 dataset are also modified by artificially included various rates (3%, 5%, 7% and 10%) of outliers to investigate the performance of the proposed method. Outliers are distributed randomly and it follows $N(5*\mu_i ,\sigma_i^2 )$; where $\mu_i$ and $\sigma_i^2$ were the *i*-th metabolite's mean and variance.

## 3. Results

To exhibit the performance of the proposed missing imputation technique compared to the other conventional missing imputation tools (zero, mean, median, half of the minimum value, kNN, BPCA, PPCA, EM-PCA and RF imputations), we analyzed both artificial and experimentally measured metabolomics datasets.

**Artificial Data Analysis Results**

In simulation studies, firstly, we measured the performance of the proposed missing imputation technique compared to the other existing nine missing imputation methods (zero, mean, median, half of the minimum value, kNN, BPCA, PPCA, EM-PCA and RF imputations) using the distance based measurement. Therefore, we computed the MSE between original simulated dataset and the reconstructed missing imputed dataset in both inexistence and existence of outliers. Since, we generated three types of simulated metabolomics datasets and 100 datasets for each type, therefore, we calculated the average MSE from 100 MSEs for each type of dataset in different rates of outliers (0%, 3%, 5%, 7% and 10%) and different rates (5%, 10%, 15% and 20%) of missing values . The results of above calculation were given in Figure 1, Figure 2, and Figure 3. In Figure 1, Figure 2, and Figure 3, the proposed missing value imputation technique produced lower average MSE for various rates (0%, 3%, 5%, 7% and 10%) of outliers as well as for various rates (5%, 10%, 15% and 20%) of missing values. Therefore, our developed missing imputation method is comparatively better than the other existing techniques.

Secondly, we evaluated the performance of our developed KMI method using MER (misclassification error rate), ROC (receiver operating characteristic curve) and AUC (area under the ROC curve) through DE metabolites identification for two groups and three groups datasets. To calculate the performance indices (MER, ROC curve and AUC values), we identified the DE metabolites from the different reconstructed datasets (missing were imputed by different methods) using *t*-test for two class level dataset and ANOVA for multiclass level dataset. Since, in simulated dataset the DE and EE metabolites were known, therefore, we computed the MER, ROC curve and AUC for different missing imputed datasets in both absence and presence of various rates of outliers. The above calculation procedures were given in Figure 4.

The ROC curve of DE calculation for two class and three class simulated datasets with 5% missing and various rates of outliers were depicted in Figure 5 and Figure 6 respectively. Similarly, for 10%, 15% and 20% missing values, the ROC curve were given in the supplementary material (Figure S1-Figure S6). Also Table 1 and Table 2 represented the MER and AUC values of DE calculation for two class and three class simulated datasets respectively with 5% missing data and various rates of outliers. Similarly, for 10%, 15% and 20% missing values, the MER and AUC values of DE calculation were also given in the supplementary material (Table S1-Table S6). The results of the performance measures from Figure 5, Figure 6, Table 1, Table 2, Figure S1-Figure S6 and Table S1-Table S6 showed that the proposed missing imputation method produced lower average MER, higher average AUC values for different rates (5%, 10%, 15% and 20%) of missing values and various rates (0%, 3%, 5%, 7% and 10%) of outliers. Therefore, we could say that the proposed KMI technique is comparatively better than other existing nine missing value imputation techniques.

Finally, we measured the performance of our proposed KMI technique through sample classification using only DE metabolites. The performance measures calculation procedure of different imputation methods on the basis of sample classification (using support vector machine classifier) were given in Figure 7. The ROC curve on the basis of sample classification using test dataset for two class simulated datasets with 5% and 10% missing values and various rates (3%, 5%, 7% and 10%) of outliers were depicted in Figure 8 and Figure 9 respectively. Figure 8 and Figure 9 showed that our proposed KMI technique gave higher average true positive rate (TPR) at any point of average false positive rate (FPR) compared to the others missing imputation methods in existence of different rates of outliers (3%, 5%, 7% and 10%). We also computed average MER and AUC in appearance of various rates of missing data and different percent of outliers using two and three class level data (Table 3 - Table 6). Table 3 - Table 6 showed that the proposed KMI technique produced lower average MER and higher average AUC values in various rates of missing values and different rates of outliers for two and three class level simulated metabolomics data. Therefore, in simulation studies, we could say that our proposed KMI technique is comparatively better than the other nine existing missing value imputation methods.

## Real Data Analysis Results

Here, we used four real metabolomics datasets for evaluating the efficiency of our newly developed KMI technique compared to the other missing imputation methods for real data analysis. Since, human cachexia dataset (available in R-specmine library) and treated dataset (available in R-metabolomics library) are fully defined, therefore, to explore the performance of our proposed technique, we artificially incorporated various percentage of missing values (5%, 10%, 15% and 20%) and reconstructed the data matrix using several missing value imputation methods including the proposed one. We measured the MSE between the original datasets and reconstructed datasets. We also repeated the aforesaid calculation 100 times and computed the average MSE for different rates of missing values that were presented in Figure 10. Figure 10 showed that the proposed missing value imputation technique produced lower average MSE for different rates of missing values of human cachexia dataset (Figure 10[a]) and treated dataset (Figure 10[b]) respectively. Therefore, we could say that our proposed imputation method is comparatively better performer than the other nine existing missing value imputation methods.

We also measured the competency of our proposed KMI technique using MER and AUC of sample classification for both two class hepatocellular carcinoma dataset and three classes MDA-MB-231 dataset. To evaluate the performance of all well-known missing value imputation methods in existence of outliers, we modified both the dataset by artificially incorporated different rates of outliers (3%, 5%, 7% and 10%). The performance measures calculation procedure for different missing imputation techniques were depicted in Figure 11. The calculation of performance measures (MER and AUC) using hepatocellular carcinoma dataset and MDA-MB-231 dataset were illustrated in Table 7 and Table 8 respectively. From Table 7 and Table 8, we could observe that our proposed KMI technique produced lower average MER and higher AUC values compared to other missing imputation methods in appearance of various rates of outliers. Therefore, both the simulation studies and real data analysis showed that our proposed missing value imputation method is comparatively better than the other nine existing missing value imputation methods.

## 4. Discussion

Here, we examined the performance of each missing imputation technique by optimizing the parameter settings using trial and error basis to avoid biased comparison. For example, in case of kNN imputation, we chose that *k*, for which the MSE, MER are smaller and accuracy is maximum. Since the performance of different missing imputation techniques may depends on the structure and the value/intensity of data, therefore, in this paper, we generated three types of simulated metabolomics datasets and 100 datasets for each type, and we calculated the average MSE from 100 MSEs for each type of dataset in different rates of outliers (0%, 3%, 5%, 7% and 10%) and different rates (5%, 10%, 15% and 20%) of missing values.

MAR may occur at any position of the data matrix; therefore, we generated 100 modified real dataset including different MAR position of the data matrix to measure the performance of different missing imputation techniques. To compute the performance of various missing imputation methods through MER and AUC using classification technique, we divided the dataset into two part, test dataset and training dataset. To reduce the sampling error during the calculation MER and AUC, we generated 100 training dataset and 100 test dataset for each case and computed the average MER and AUC for measuring the performance of different missing imputation methods. The detail calculation procedure of different performance measures calculated by different missing imputation methods for artificial dataset and experimentally measured (i.e., real ) datasets were shown in Figure 4 & Figure 7 and Figure 11 respectively. The url of the R package and user manual is https://github.com/NishithPaul/tWLSA .

## Conclusion

Selection of missing imputation method plays a vital role for consecutive metabolomics data analysis. However, all the conventional missing value imputation methods are more or less affected by outlying observations. Thus, in this paper, we have developed a new outlier-robust kernel weight based two-way alternating weighted least square approach for imputing missing values. We also measured the performance of our newly developed KMI technique compared to the other nine existing methods (zero, mean, median, half of the minimum value, kNN, BPCA, PPCA, EM-PCA and RF imputations) through both artificial and real metabolomics data analysis. Observing the computational results, we could conclude that our developed missing value imputation method is comparatively better than the conventional nine missing value imputation methods in both appearance and non-appearance of outliers. For this reason, our suggestion is to apply our proposed two-way kernel weighted least square based missing value imputation method instead of conventional missing imputation methods to substitute the missing values in metabolomics datasets for consecutive univariate, multivariate as well as exploratory metabolomics data analysis.

### Acknowledgment

### Author contributions statement

Nishith Kumar (NK) worked to develop the two-way kernel weighted least square based missing imputation technique. NK also analyzed the data, drafted the manuscript, and executed the statistical analysis. Md. Aminul Hoque (MAH) and Masahiro Sugimoto (MS) coordinated and supervised the project. All authors carefully read and finally approved the manuscript

### Additional Information

The authors declare no competing financial interests.

## References

1. Gromski, P.S., Xu, Y., Kotze, H.L., Correa, E., Ellis, D.I., Armitage, E.G., Turner, M.L. & Goodacre, R. Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites* **4**, 433-452 (2014).

2. Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T., & Ni, Y. Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci. Rep* **8**, 1-10 (2018).
3. Hrydziuszko, O., & Viant, M. R. Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics* **8**, 161-174 (2012).
4. Steuer, R., Morgenthal, K., Weckwerth, W., & Selbig, J. A gentle guide to the analysis of metabolomic data. *Methods Mol. Biol.* **358**, 105–126 (2007).
5. Di Guida, R., Engel, J., Allwood, J.W., Weber, R.J., Jones, M.R., Sommer, U., Viant, M.R., & Dunn, W.B. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics* **12**, 93 (2016).
6. Armitage, E. G., Godzien, J., Alonso-Herranz, V., López-Gonzálvez, Á., & Barbas, C. Missing value imputation strategies for metabolomics data. *Electrophoresis*, **36**, 3050-3060 (2015).
7. Navarrete, A., Armitage, E.G., Musteanu, M., García, A., Mastrangelo, A., Bujak, R., López-Casas, P.P., Hidalgo, M. & Barbas, C. Metabolomic evaluation of Mitomycin C and rapamycin in a personalized treatment of pancreatic cancer. *Pharmacol. Res. Persp* **2**, e00067 (2014).
8. Qiu, Y., Rajagopalan, D., Connor, S.C., Damian, D., Zhu, L., Handzel, A., Hu, G., Amanullah, A., Bao, S., Woody, N. & MacLean, D. Multivariate classification analysis of metabolomic data for candidate biomarker discovery in type 2 diabetes mellitus. *Metabolomics* **4**, 337-346 (2008).
9. Kirwan, J. A., Weber, R. J., Broadhurst, D. I., & Viant, M. R. Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. *Sci. Data* **1**, 1-13 (2014).
10. Krug, S., Kastenmüller, G., Stückler, F., Rist, M.J., Skurk, T., Sailer, M., Raffler, J., Römisch-Margl, W., Adamski, J., Prehn, C. & Frank, T. The dynamic range of the human metabolome revealed by challenges. *FaSEB J.* **26**, 2607-2619 (2012).
11. Sun, X., & Weckwerth, W. COVAIN: a toolbox for uni-and multivariate statistics, time-series and correlation network analysis and inverse estimation of the differential Jacobian from metabolomics covariance data. *Metabolomics* **8**, 81-93 (2012).
12. Madhu, G., Bharadwaj, B. L., Vardhan, K. S., & Chandrika, G. N. A Normalized Mean Algorithm for Imputation of Missing Data Values in Medical Databases. In *Innovations in Electronics and Communication Engineering*, 773-781 (Springer, Singapore, 2020).

13. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525 (2001).
14. Nyamundanda, G., Brennan, L., & Gormley, I. C. Probabilistic principal component analysis for metabolomic data. *BMC Bioinform.* **11**, 571 (2010).
15. Xia, J., & Wishart, D. S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat. Protoc.* **6**, 743–760 (2011).
16. Ilin, A., & Raiko, T. Practical approaches to principal component analysis in the presence of missing values. *J. Mach. Learn. Res.* **11**, 1957-2000 (2010).
17. Jansen, J. J., Hoefsloot, H. C., Boelens, H. F., Van Der Greef, J., & Smilde, A. K. Analysis of longitudinal metabolomics data. *Bioinformatics* **20**, 2438-2446 (2004).
18. Lin, T. H. A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Qual. Quant.* **44**, 277-287 (2010).
19. Roweis, S. T. EM algorithms for PCA and SPCA. In *Advances in neural information processing systems*, 626-632 (1998).
20. Stekhoven, D. J., & Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, **28**, 112-118 (2012).
21. Pedreschi, R., Hertog, M.L., Carpentier, S.C., Lammertyn, J., Robben, J., Noben, J.P., Panis, B., Swennen, R. and Nicolaï, B.M. Treatment of missing values for multivariate statistical analysis of gel-based proteomics data. *Proteomics* **8**, 1371-1383 (2008).
22. Scheel, I., Aldrin, M., Glad, I., Sorum, R., Lyng, H., & Frigessi, A. The influence of missing values imputation on detection of differentially expressed genes from microarray data. *Bioinformatics* **21**, 4272–4279 (2005).
23. Brevern, A. G., Hazout, S., & Malpertuy, A. Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinform.* **5**, 114 (2004).
24. Blanchet, L., & Smolinska, A. Data fusion in metabolomics and proteomics for biomarker discovery. In *Statistical Analysis in Proteomics,* 209-223 (Humana Press, 2016).
25. Tzoulaki, I., Ebbels, T. M., Valdes, A., Elliott, P., & Ioannidis, J. P. Design and analysis of metabolomics studies in epidemiologic research: a primer on-omic technologies. *Am. J. Epidemiol* **180**, 129-139 (2014).
26. Tibshirani, R., & Hastie, T. Outlier sums for differential gene expression analysis. *Biostatistics* **8,** 2-8 (2007).

27. Kumar, N., Hoque, M. A., Shahjaman, M., Islam, S. M., & Mollah, M. N. H. Metabolomic Biomarker Identification in Presence of Outliers and Missing Values. *BioMed Res. Int* **2017**, (2017).
28. Kotze, H.L., Armitage, E.G., Sharkey, K.J., Allwood, J.W., Dunn, W.B., Williams, K.J. & Goodacre, R. A novel untargeted metabolomics correlation-based network analysis incorporating human metabolic reconstructions. BMC Syst. Biol **7**, 107 (2013).

## Figure legends

**Figure 1.** Performance investigation of different missing imputation techniques using average MSE for without class level data.

**Figure 2.** Performance investigation of different missing imputation techniques using average MSE for two class level data.

**Figure 3.** Performance investigation of different missing imputation techniques using average MSE for three class level data.

**Figure 4.** Performance measures calculation procedure on the basis of DE calculation.

**Figure 5.** Performance investigation of different missing value imputation techniques using ROC curve of DE calculation for two class level dataset with 5% missing values in absence and presence of outliers.

**Figure 6.** Performance investigation of different missing value imputation techniques using ROC curve of DE calculation for three class level dataset with 5% missing values in absence and presence of outliers.

**Figure 7.** Performance measures calculation procedure on the basis of sample classification.

**Figure 8.** Performance investigation of different missing value imputation techniques using ROC curve of sample classification for two class level dataset with 5% missing values in presence of outliers.

**Figure 9.** Performance investigation of different missing value imputation techniques using ROC curve of sample classification for two class level dataset with 10% missing values in presence of outliers.

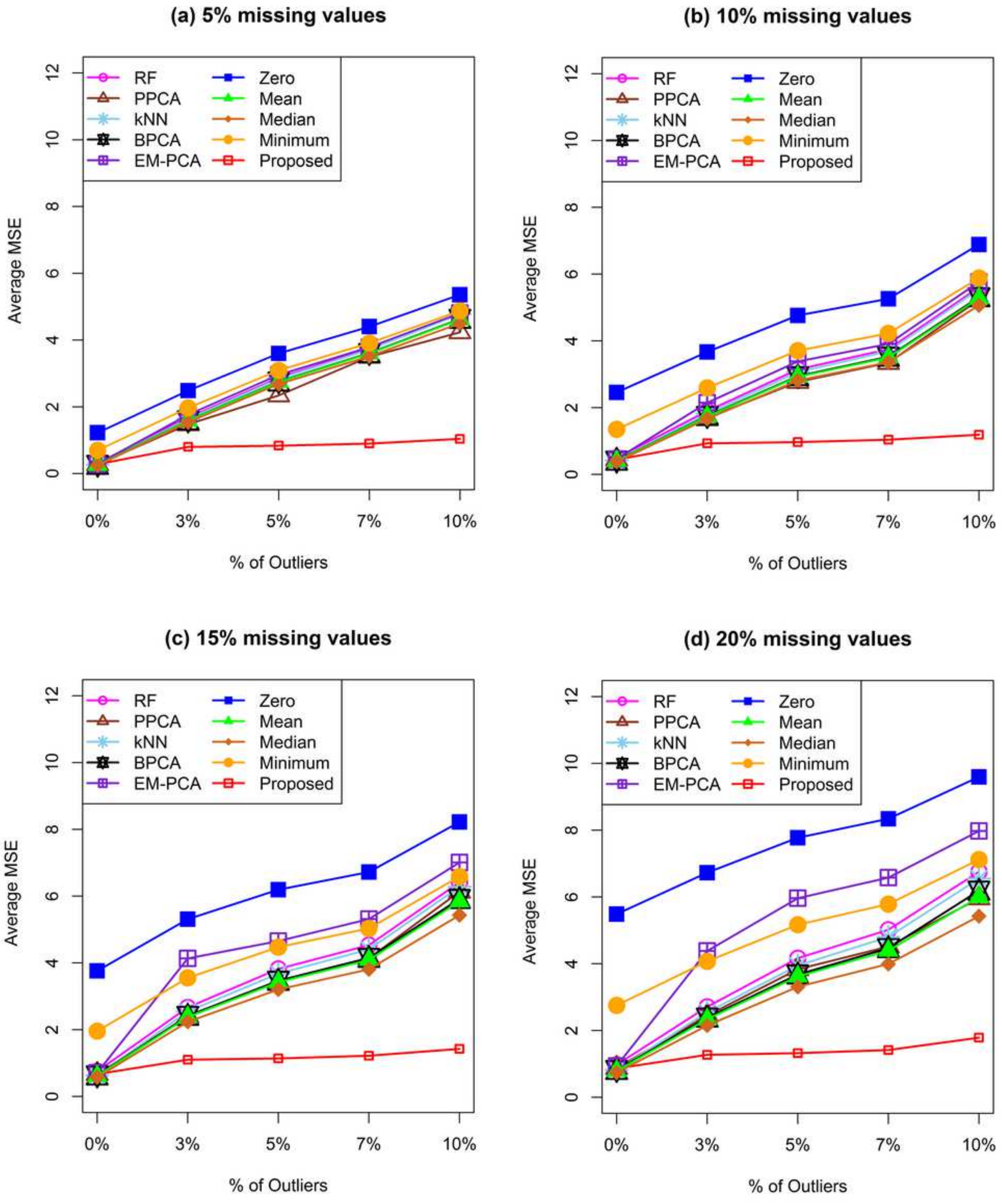
**Figure 10.** Performance investigation of different missing value imputation techniques using MSE calculation for different rates of missing values of (a) human cachexia dataset and (b) treated dataset.

**Figure 11.** Performance measures calculation procedure for real dataset on the basis of sample classification.

## Diagram Legend

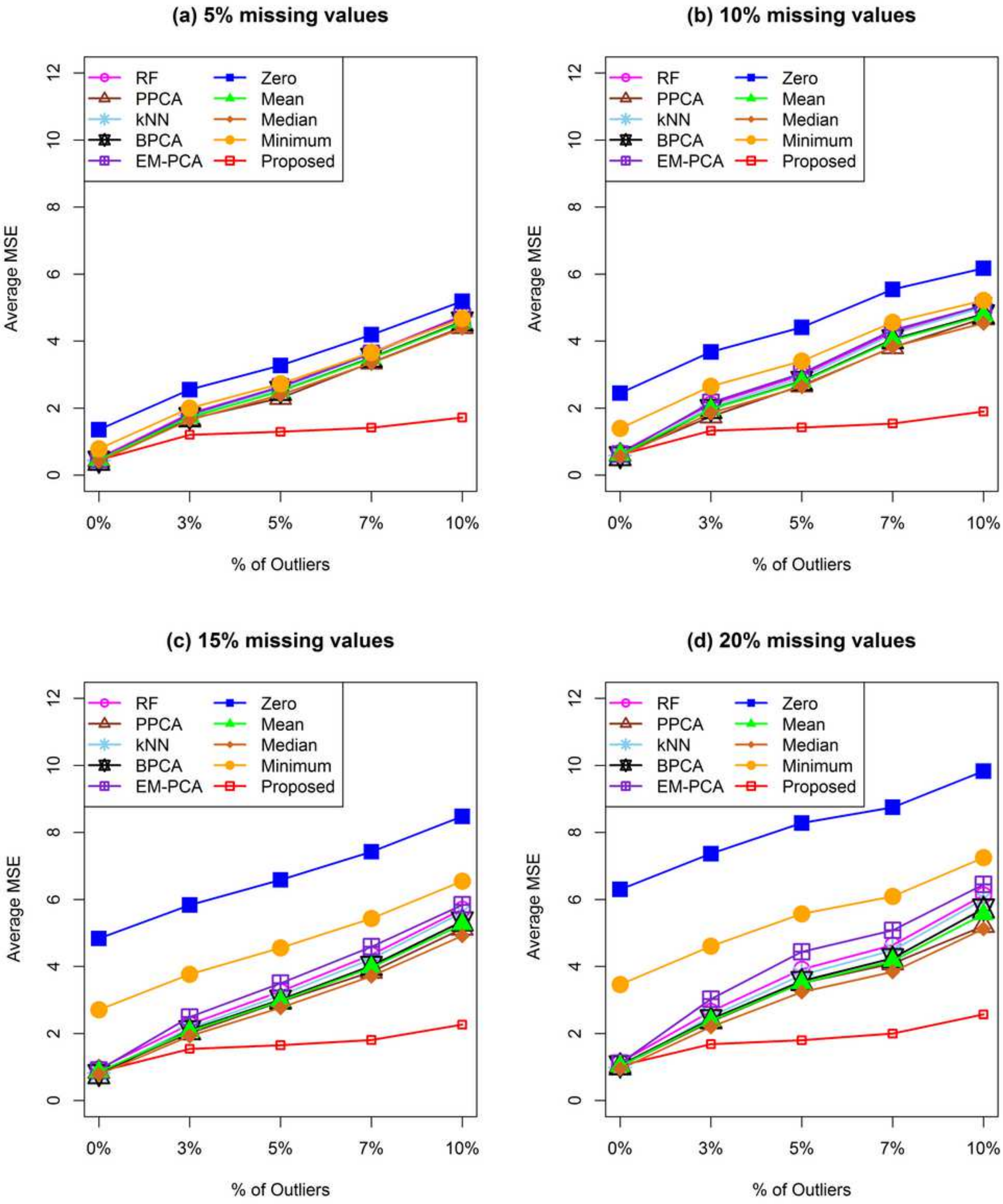
**Diagram 1.** $\lambda$ selection procedure.

# Figures



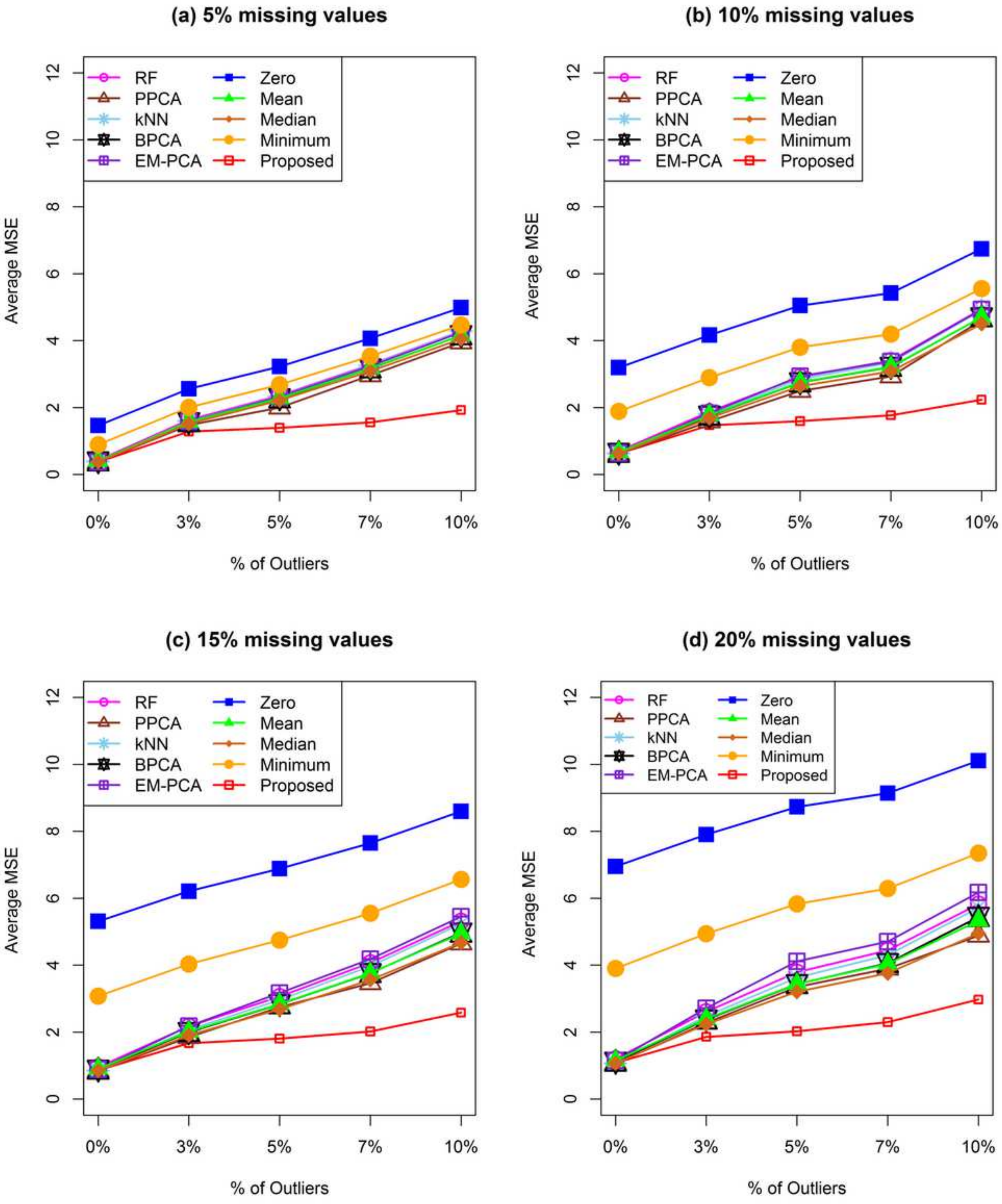
## Figure 1

Performance investigation of different missing imputation techniques using average MSE for without class level data.



## Figure 2

Performance investigation of different missing imputation techniques using average MSE for two class level data.



## Figure 3

Performance investigation of different missing imputation techniques using average MSE for three class level data.

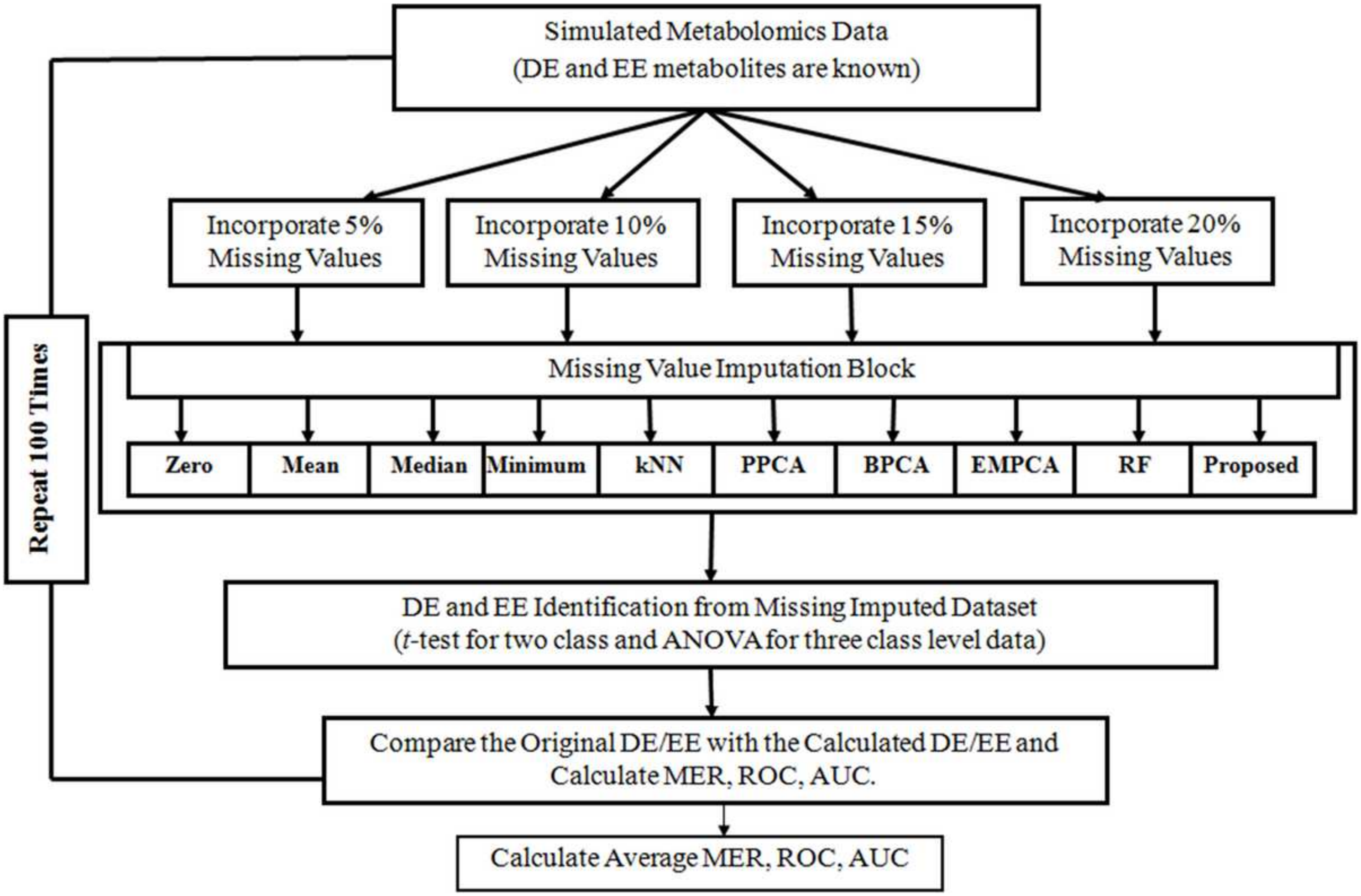
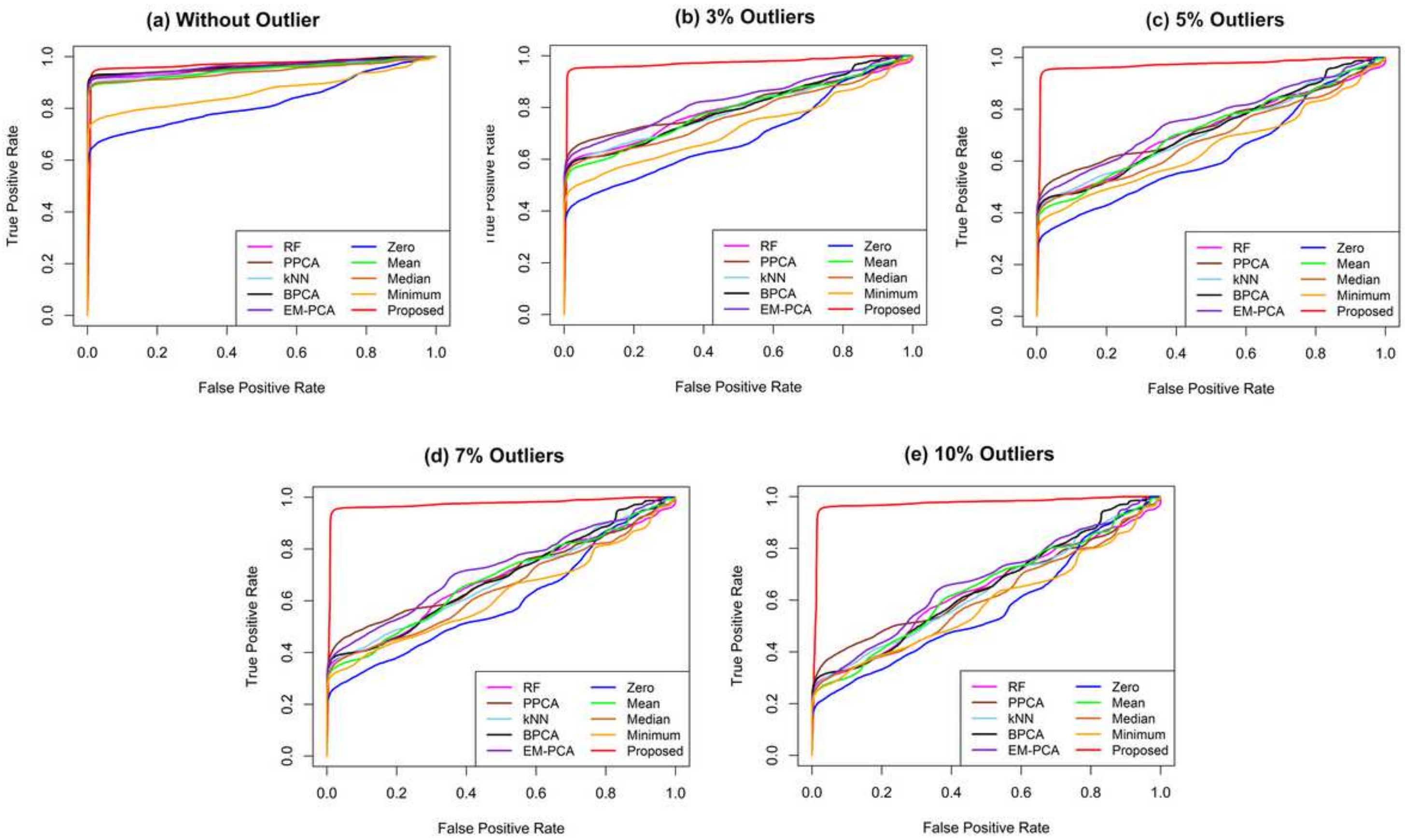

Simulated Metabolomics Data
(DE and EE metabolites are known)
Incorporate 5% Missing Values
Incorporate 10% Missing Values
Incorporate 15% Missing Values
Incorporate 20% Missing Values
Repeat 100 Times
Missing Value Imputation Block
Zero
Mean
Median
Minimum
kNN
PPCA
BPCA
EMPCA
RF
Proposed
DE and EE Identification from Missing Imputed Dataset
(t-test for two class and ANOVA for three class level data)
Compare the Original DE/EE with the Calculated DE/EE and Calculate MER, ROC, AUC.
Calculate Average MER, ROC, AUC

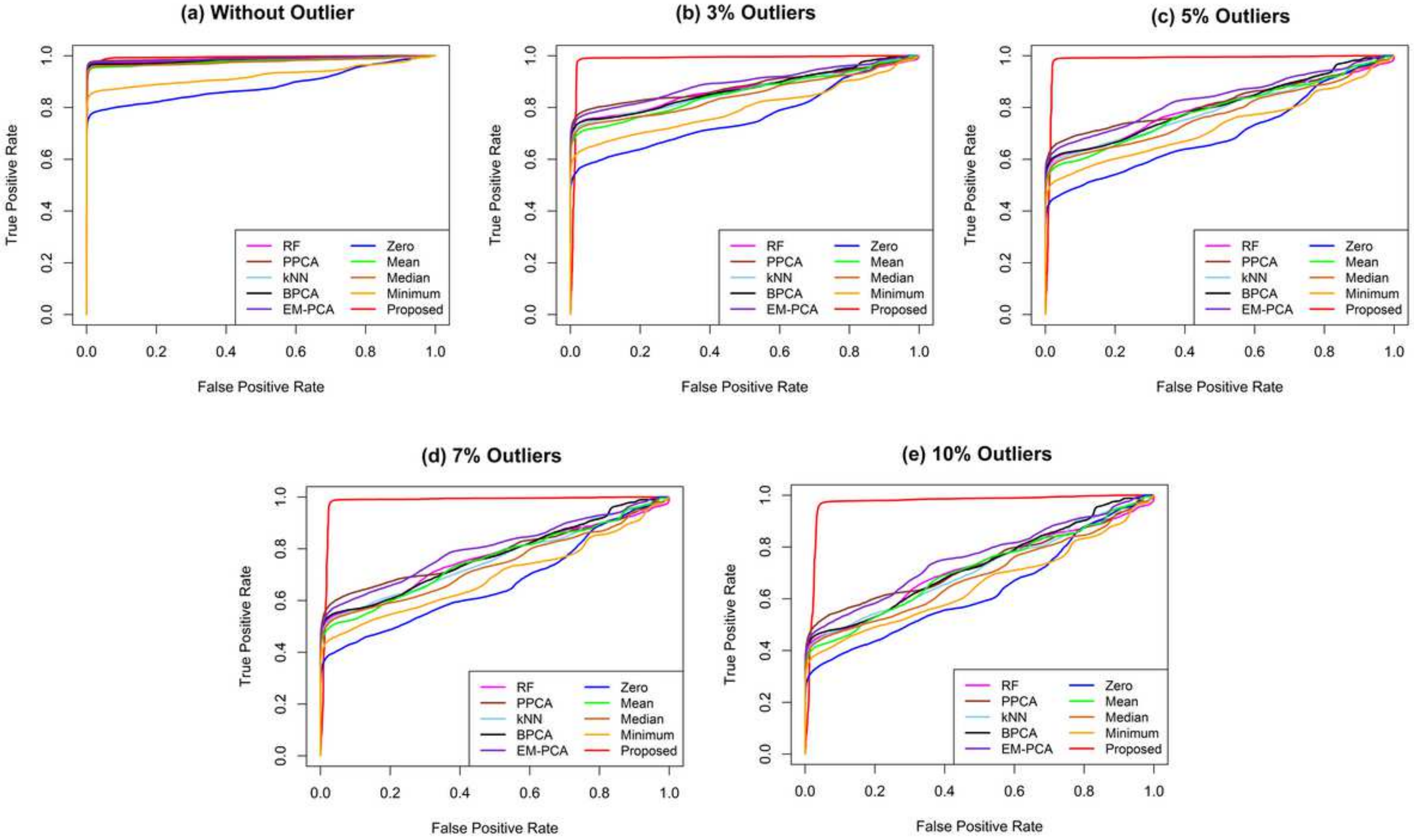
Figure 4

Performance measures calculation procedure on the basis of DE calculation.



## Figure 5

Performance investigation of different missing value imputation techniques using ROC curve of DE calculation for two class level dataset with 5% missing values in absence and presence of outliers.



## Figure 6

Performance investigation of different missing value imputation techniques using ROC curve of DE calculation for three class level dataset with 5% missing values in absence and presence of outliers.

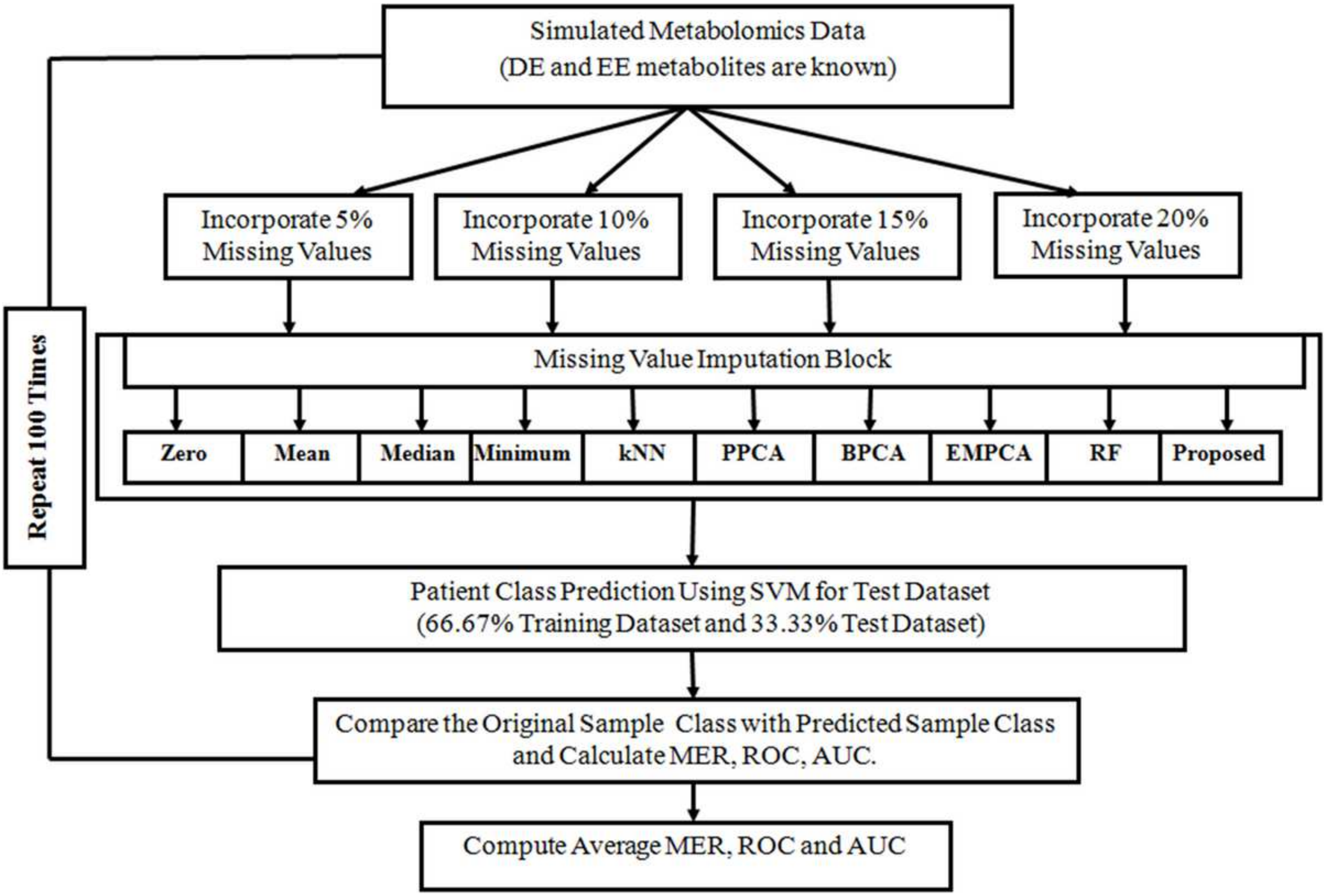
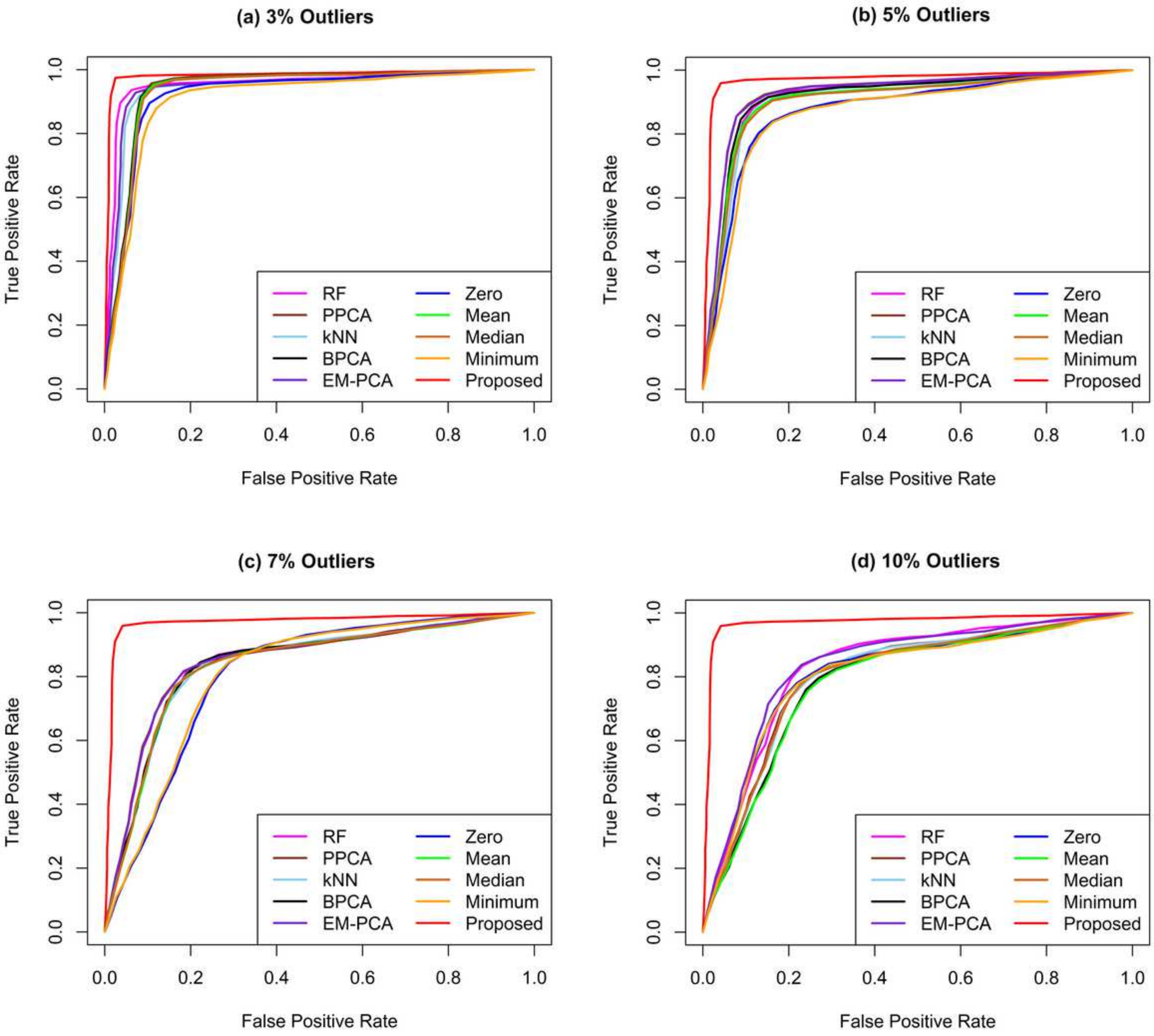

Simulated Metabolomics Data
(DE and EE metabolites are known)
Incorporate 5% Missing Values
Incorporate 10% Missing Values
Incorporate 15% Missing Values
Incorporate 20% Missing Values
Missing Value Imputation Block
Zero
Mean
Median
Minimum
kNN
PPCA
BPCA
EMPCA
RF
Proposed
Repeat 100 Times
Patient Class Prediction Using SVM for Test Dataset
(66.67% Training Dataset and 33.33% Test Dataset)
Compare the Original Sample Class with Predicted Sample Class and Calculate MER, ROC, AUC.
Compute Average MER, ROC and AUC

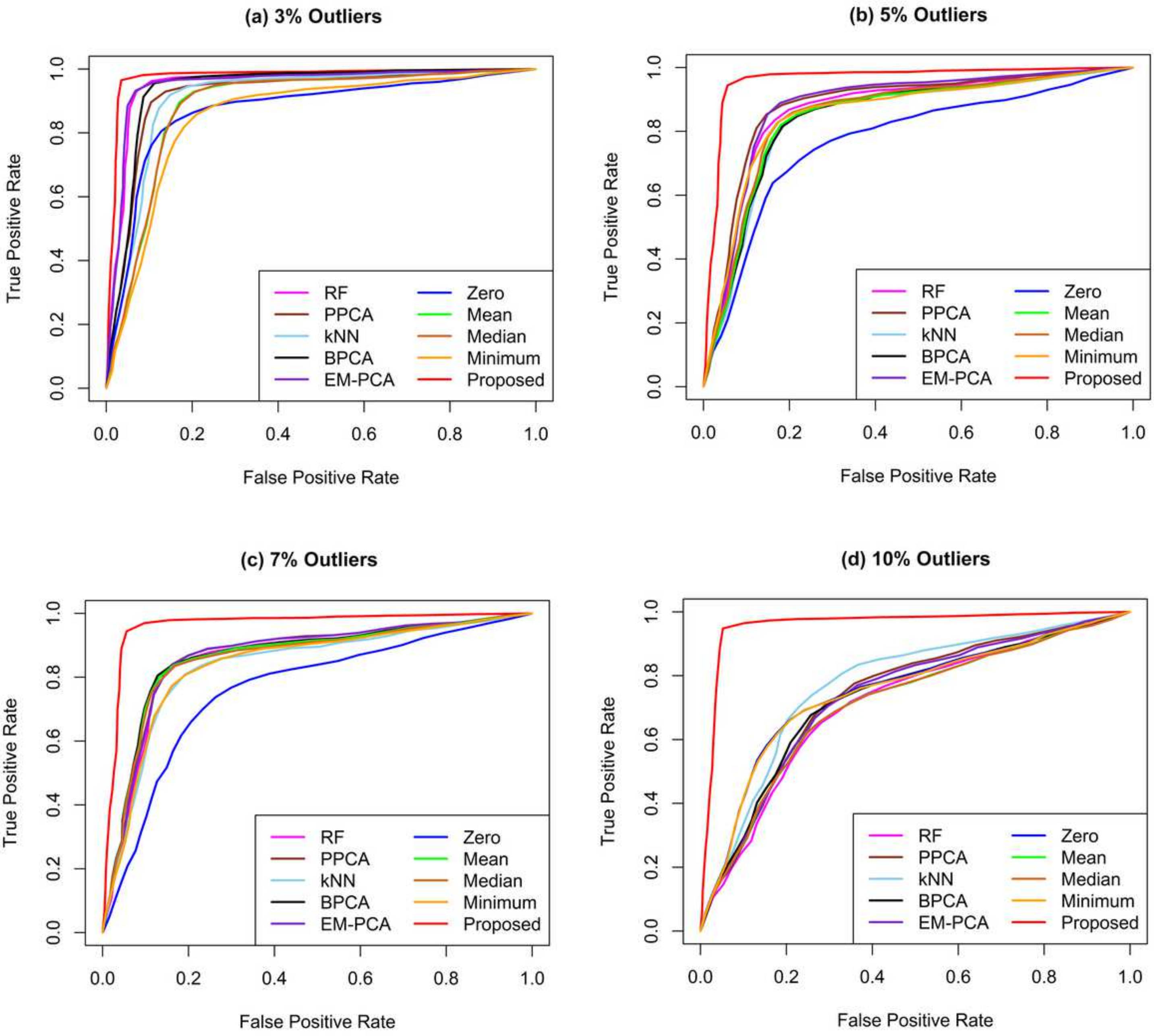
Figure 7

Performance measures calculation procedure on the basis of sample classification.



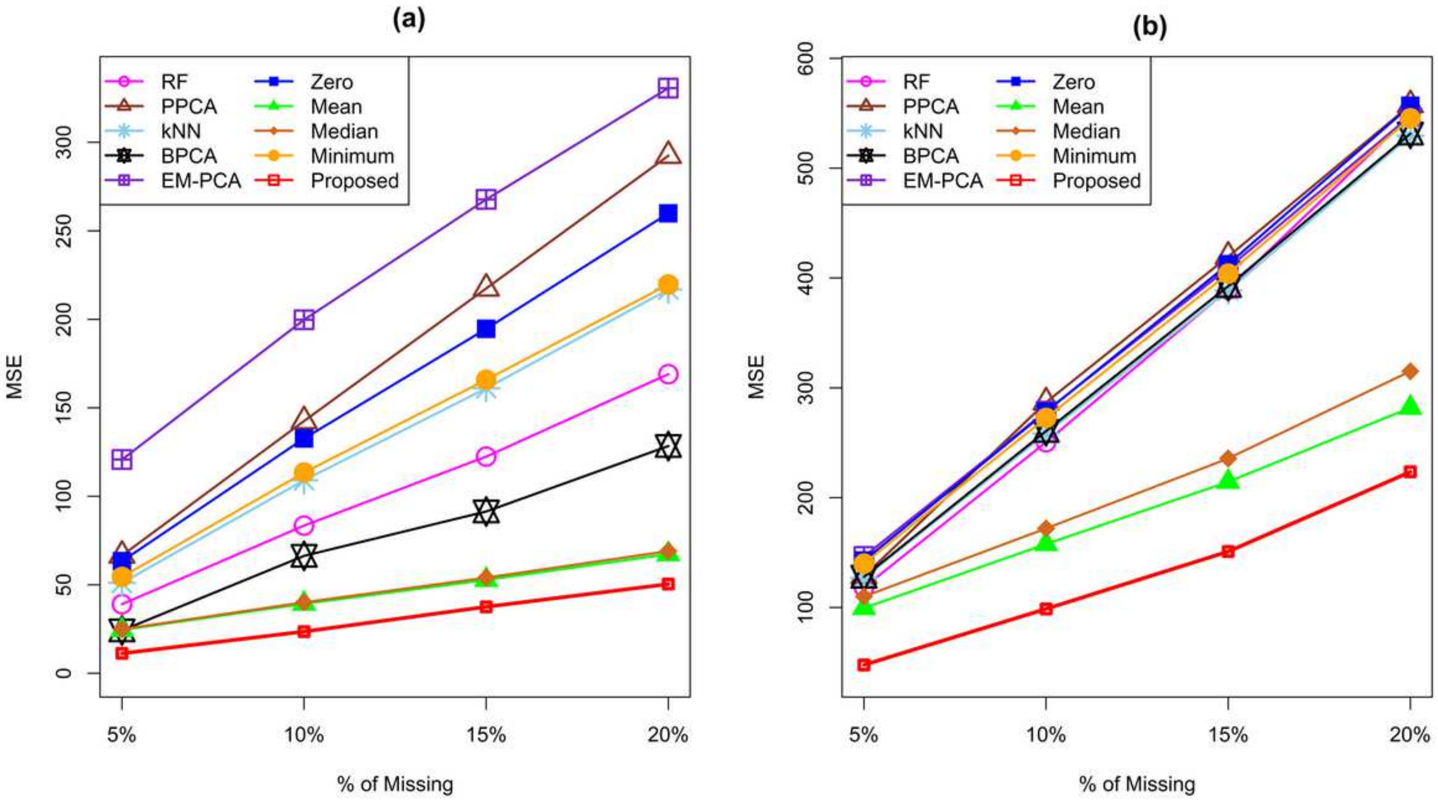
## Figure 8

Performance investigation of different missing value imputation techniques using ROC curve of sample classification for two class level dataset with 5% missing values in presence of outliers.



## Figure 9

Performance investigation of different missing value imputation techniques using ROC curve of sample classification for two class level dataset with 10% missing values in presence of outliers.



## Figure 10

Performance investigation of different missing value imputation techniques using MSE calculation for different rates of missing values of (a) human cachexia dataset and (b) treated dataset.

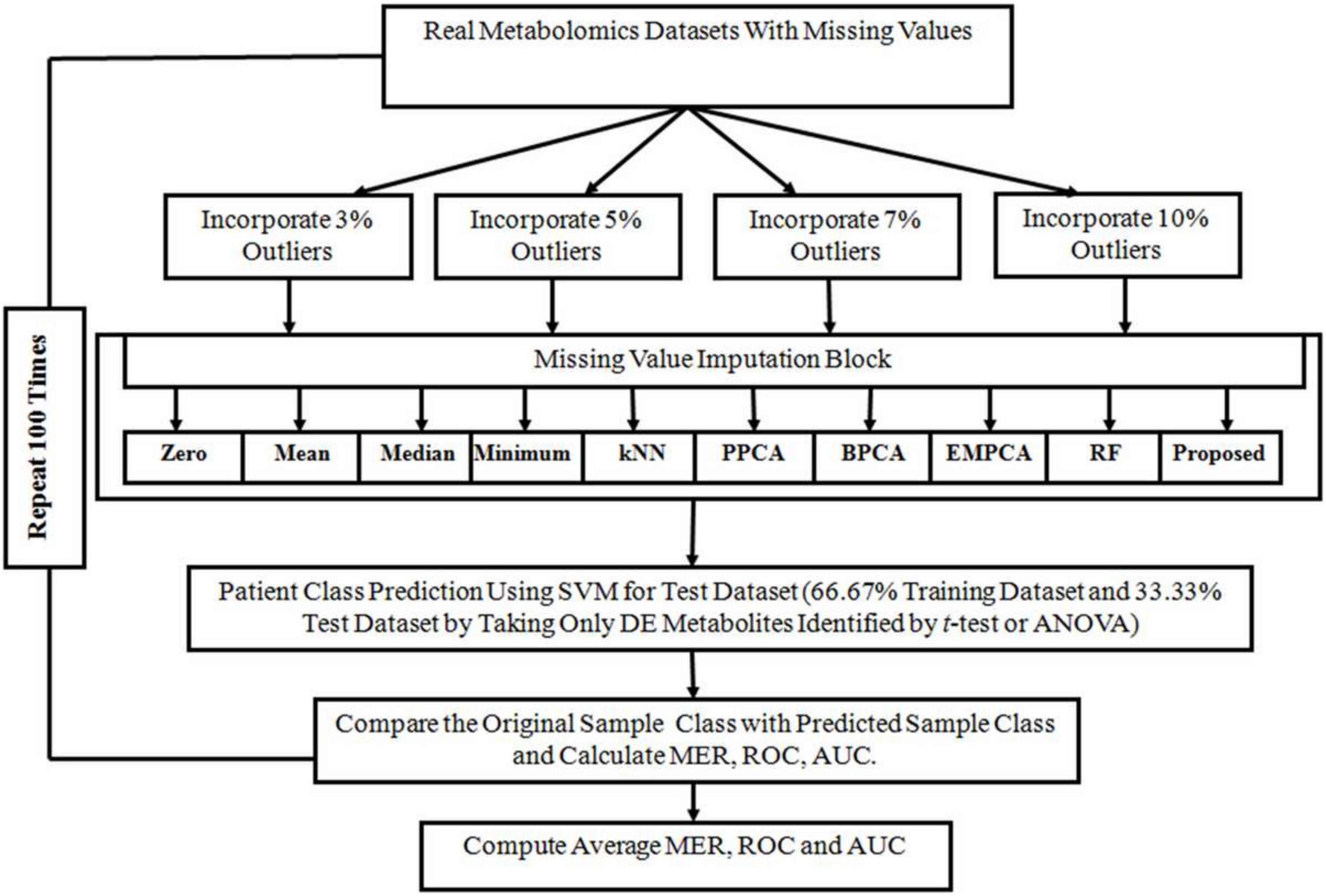
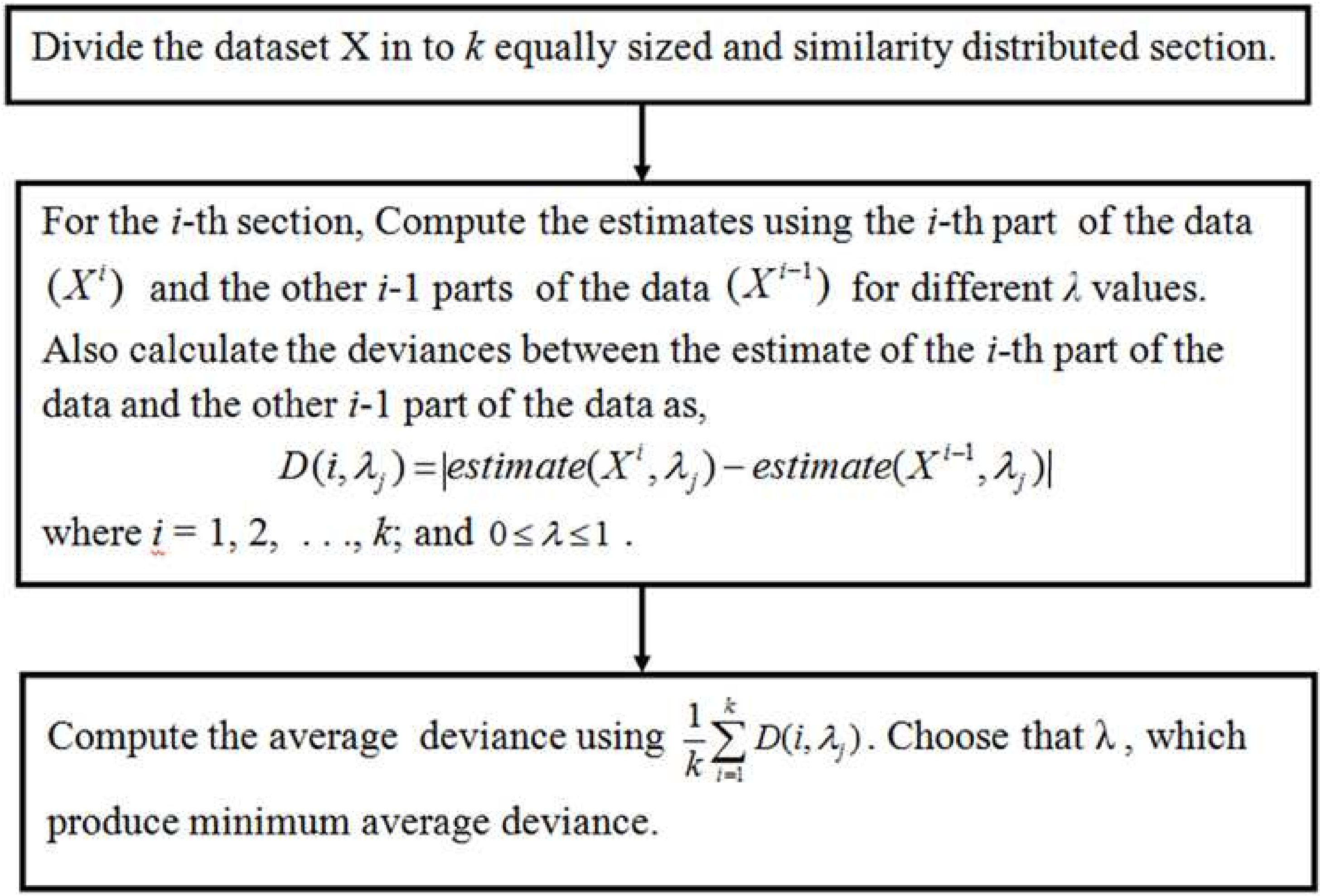

Real Metabolomics Datasets With Missing Values
Incorporate 3% Outliers
Incorporate 5% Outliers
Incorporate 7% Outliers
Incorporate 10% Outliers
Repeat 100 Times
Missing Value Imputation Block
Zero
Mean
Median
Minimum
kNN
PPCA
BPCA
EMPCA
RF
Proposed
Patient Class Prediction Using SVM for Test Dataset (66.67% Training Dataset and 33.33% Test Dataset by Taking Only DE Metabolites Identified by t-test or ANOVA)
Compare the Original Sample Class with Predicted Sample Class and Calculate MER, ROC, AUC.
Compute Average MER, ROC and AUC

Figure 11

Performance measures calculation procedure for real dataset on the basis of sample classification.



Divide the dataset X in to $k$ equally sized and similarity distributed section.

For the $i$-th section, Compute the estimates using the $i$-th part of the data $(X^i)$ and the other $i$-1 parts of the data $(X^{i-1})$ for different $\lambda$ values. Also calculate the deviances between the estimate of the $i$-th part of the data and the other $i$-1 part of the data as,

$$D(i,\lambda_j) = |estimate(X^i,\lambda_j) - estimate(X^{i-1},\lambda_j)|$$

where $i$ = 1, 2, . . ., $k$; and $0 \le \lambda \le 1$ .

Compute the average deviance using $\frac{1}{k}\sum_{i=1}^{k} D(i,\lambda_j)$. Choose that $\lambda$, which produce minimum average deviance.

**Diagram 1.** $\lambda$ selection procedure.

**Figure 12**

λ selection procedure.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TablesV1.doc
- SupplementaryInformationV1.doc