How to cite this article:

Kumaran, S. R., Othman, M. S., & Yusuf, L. M. (2020). Estimation of missing values using optimized hybrid fuzzy c-means and majority vote for microarray data. *Journal of Information and Communication Technology, 19*(4), 459-482. https://doi.org/10.32890/jict2020.19.4.1

ESTIMATION OF MISSING VALUES USING OPTIMISED HYBRID FUZZY C-MEANS AND MAJORITY VOTE FOR MICROARRAY DATA

Shamini Raja Kumaran, Mohd Shahizan Othman &
Lizawati Mi Yusuf

School of Computing,
Universiti Teknologi Malaysia, Malaysia

shamini.rajakumaran@hotmail.com; shahizan, lizawati@utm.my

ABSTRACT

Missing values are a huge constraint in microarray technologies towards improving and identifying disease-causing genes. Estimating missing values is an undeniable scenario faced by field experts. The imputation method is an effective way to impute the proper values to proceed with the next process in microarray technology. Missing value imputation methods may increase the classification accuracy. Although these methods might predict the values, classification accuracy rates prove the ability of the methods to identify the missing values in gene expression data. In this study, a novel method, Optimised Hybrid of Fuzzy C-Means and Majority Vote (*opt*-FCMMV), was proposed to identify the missing values in the data. Using the Majority Vote (MV) and optimisation through Particle Swarm Optimisation (PSO), this study predicted missing values in the data to form more informative and solid data. In order to verify the effectiveness of *opt*-FCMMV, several experiments were carried out on two publicly available microarray datasets (i.e. Ovary and Lung Cancer) under three missing value mechanisms with five different percentage values in the biomedical domain using Support Vector

Received: 10/7/2019 **Revised:** 14/10/2019 **Accepted:** 18/2/2020 **Published:** 20/8/2020

Machine (SVM) classifier. The experimental results showed that the proposed method functioned efficiently by showcasing the highest accuracy rate as compared to the one without imputations, with imputation by Fuzzy C-Means (FCM), and imputation by Fuzzy C-Means with Majority Vote (FCMMV). For example, the accuracy rates for Ovary Cancer data with 5% missing values were 64.0% for no imputation, 81.8% (FCM), 90.0% (FCMMV), and 93.7% (*opt*-FCMMV). Such an outcome indicates that the *opt*-FCMMV may also be applied in different domains in order to prepare the dataset for various data mining tasks.

Keywords: Fuzzy C-means, majority vote**,** missing values, microarray data, data optimisation.

INTRODUCTION

In many areas, the quality of data is a very serious problem in the current rapid world that produces millions of data each day that are often noisy and incomplete. Nevertheless, the issues from missing data are ubiquitous in the healthcare sector especially in microarray experiments that are able to generate thousands of gene expression datasets with missing expression values. The consequences faced by real-world healthcare research centres, such as the production of biased data and invalid inferences, undermine the purpose of data (Suphanchaimat et al., 2017). This is due to experimental errors, insufficient resolutions, and scratches or dust in slides during the laboratory processes (Yaraghi et al., 2012). As mentioned by Ouyang et al. (2004), every microarray experiment virtually contains missing expressions, and this affects more than 90% of the genes. During these scenarios, the extracted gene expression microarray datasets are unable to guarantee complete and useful knowledge that may influence the validity of the data. Meanwhile, the fundamental goal of microarray data is to detect the expressions of thousands of genes, identify disease-causing genes (Pino Angulo et al., 2018), accelerate molecular biology experiments, and find the functions of genes, genetic networks, and biomarker genes (Li et al., 2010). Therefore, it is important to consider the treatment of missing values before analysing the microarray data.

There are existing missing value strategies that have been developed and deployed in the gene expression data to promote data quality and reliability. The common treatment of missing values for microarray data is classified into three categories. Ignorance is the simplest solution to delete the records of data with missing values using listwise and pairwise deletion methods. Nevertheless, these deletion methods might drop abundant values in one process and reduce the accuracy rate in order to identify the disease-causing genes. The second category, tolerance, discards missing points in the data. Even though this is a low-cost solution, it might produce low-quality datasets. The third category, imputation, is one of the best methods that can renew the whole dataset in order to prove the best means to process the missing values in the experiments (Tian et al., 2012; Hourani & Emary, 2009). Accordingly, the imputation method attempts to increase the relevancy and knowledge from the data that are able to construct a complete dataset. Taking all into account, a new imputation method was proposed to impute the missing values based on existing values in the data that are able to construct more information and knowledge. Practically, the proposed method is realised as a hybrid of Fuzzy C-Means (FCM), Majority Vote (MV), and Particle Swarm Optimisation (PSO), which is termed as *opt*-FCMMV in this study. An optimisation's contribution is to minimise and maximise the decision-making algorithm normally adapted to the approximation methods (Shehab et al., 2018). Therefore, the central idea of the imputation method is to use optimisation as the key in improvising and predicting the best missing values. In this study, *opt*-FCMMV is investigated as a solution for gene expression datasets.

An Optimised Hybrid of Fuzzy C-means and Majority Vote (*opt*-FCMMV) using PSO is proposed to impute the missing values in order to provide better information on the data. The effectiveness of the proposed *opt*-FCMMV in terms of solution quality and computational efficiency was demonstrated at various level (5%, 10%, 30%, 50%, and 80%) and missing value mechanisms such as Missing at Random (MAR), Missing Completely at Random (MCAR), and Missing Not at Random (MNAR) on two publicly available microarray data. The datasets with different missing values and mechanisms were tested with the proposed method (*opt*-FCMMV), FCM, and Fuzzy C-Means with Majority Vote (FCMMV). The performance of classification showed that the proposed method is able to produce a higher accuracy rate due to the optimisation by metaheuristic algorithms such as PSO. Considering the increasing demand of analysing data in various domains such as biomedical, this study hopes that it will be able to provide a new direction for missing value imputation by overcoming issues such as trap in local minima and high level of objective function. The remainder of the article is divided into four main parts. The upcoming section describes the theory related to missing values and presents a literature survey of the existing methods. Next, the missing value imputation method termed as the Optimised Hybrid of Fuzzy C-Means and Majority Vote (*opt*-FCMMV) is proposed. Then, the experimental results obtained from thirty datasets are presented while the conclusion of the study is presented at the end.

RELATED WORKS

In the context of missing value mechanisms, the mechanisms can be divided into three main groups: Missing at Random (MAR), Missing Completely at Random (MCAR), and Missing Not at Random (MNAR) (Suphanchaimat et al., 2017; Dibal et al., 2017; Kellermann et al., 2016; Tshering et al., 2013). MAR assesses the probability of missing data that do not depend on unobserved data; however, it also does not depend on available information. MAR consists of equal values of missing data that are randomly distributed within one or more sub-samples of data (Rubin, 1976); *P (missing| observed, unobserved) = P (missing | observed)* (Dibal et al., 2017). An example of an MAR scenario is women are more likely to get breast cancer; however, the probability of women who come for breast cancer check-up to get a diagnosis is the same for all women. In contrast, MCAR defines the probability of missing values on one variable is unrelated to other observed variables; *P (missing| observed, unobserved) = P (missing)* (Dibal et al., 2017; Tshering et al., 2013). For example, a breast cancer test has been performed on the patients; however, the mammogram is unable to function properly, whereby the results might show missing points completely at random. Meanwhile, MNAR is the probability of data that have fields of missing values and depend on the values of attributes; *P (missing| observed, unobserved)*. MNAR cannot be quantified because the missing values depend on the values (Dibal et al., 2017; Tshering et al., 2013). An example of an MNAR scenario is breast cancer patients might be required to undergo chemotherapy weekly to screen whether the cancer has grown or spread. However, if the patient fails to show up for the chemotherapy sessions, then, the missing data points are related to the unobserved spread of cancers and this is classified as MNAR. In reality, most research for microarray experiments have been devoted to MAR or MCAR mechanisms, while very few research have been conducted on MNAR scenarios. Lazar et al. (2016) are considered as one of the motivations to conduct this research. With the knowledge of missing value mechanisms, it is practical enough to identify the appropriate analysis method recommended for the datasets. In many situations, missing values are required to be imputed in order to further analyse the imputed dataset (Bertsimas et al., 2017).

There are many imputation methods proposed specifically for microarray datasets. A number of effective imputations that have been used are clustering (Salleh & Samat, 2017) and classification algorithms (Tsai et al., 2018). Most articles proposed cluster-based algorithms and utilised high dimensional microarray datasets with a large number of features and samples that might directly affect the clustering performance (Keerin et al., 2016; Chattopadhyay et al., 2015; Gupta et al., 2015; Keerin et al., 2012) Moreover, the clustering performance is highly dependent on the number of clusters and with such conditions of samples, the selection of clusters will be crucial. Therefore, the researchers must handle the selection of clusters with a more detailed analysis in the algorithm of the selection part. Paul et al. (2017) utilised a pattern similarity matching algorithm, while Baraldi et al. (2015) used fuzzy similarity to impute missing values and the optimised fuzzy rule for gene selection. Indeed, gene selection is an important phase to pre-process the data and improvise the classification performance. However, missing values in the dataset must be handled well before identifying the disease-causing genes. The main disadvantage of pattern similarity matching is due to the distance that affects the dimensions with high dissimilarity (Tung et al., 2006), which might reflect its drawback in imputing the missing values in the data.

Some articles proposed Fuzzy C-Means as the clustering algorithm handling missing values (Saha et al., 2016; Pourhasem et al., 2010). However, the main drawback of fuzzy clustering is sensitivity at the initialisation phase, which will decrease the efficiency of the method. Consequently, this is the main reason FCM is hybridised with MV in this research. Furthermore, one of the commonly used methods for missing values is k-nearest neighbour (kNN) imputation (De Silva & Perera, 2017; Suyundikov et al., 2015; Keerin et al., 2012). In this process of kNN to impute the missing values, the intra-cluster dissimilarity is measured using the summation of distances between the data. However, the drawbacks of kNN imputation are the choice of the function of distances, time-consuming due to the large database, and choice number of neighbours (Edgar & Rodirguez, 2004). Additionally, Local Least Squares Imputation (LLSimpute) is common in estimating missing values. Yu et al. (2017), Bose et al. (2013), and Qin and Lee (2010) used LLSimpute to estimate missing values in microarray gene expression data. Nevertheless, one disadvantage of LLSimpute is that the optimal number of neighbours is based on the heuristic search that might elevate the computational cost of the algorithm.

To conclude, with evidence from the recent research, this article would like to suggest that advances in optimisation have shown promising progress in machine learning to be applied in missing value situations. This idea can be used to solve the missing value issues in microarray datasets as the optimisation process is able to offer effective solutions in a difficult scenario. The ability of the optimisation method can be used to minimise the missing data error (Marwala, 2009). Despite imputing the best predicted missing values in the data, the proposed method is able to provide informative data. The proposed method in this article aims to utilise the power of the optimisation and hybrid technique in imputing the accurate values. The advantages of this new method are: a) construction of values that are more accurate; and b) use of optimisation to minimise the difference measure between clusters centres and the values that directly minimise the data error. Moreover, to handle the reliability and validity of data with high accuracy rates is an important challenge faced in missing value scenarios.

RESEARCH DESIGN

In this section, a new missing value imputation method termed as Optimised Hybrid of Fuzzy C-Means and Majority Vote (*opt*-FCMMV) is proposed for microarray datasets. Although FCM is able to impute missing values, there is room to improvise FCM. Therefore, FCM is hybridised with MV in this study. Through this, MV is able to construct many accurate values in the missing data for best selection on the estimation of missing values. After this hybridisation, FCMMV will be optimised using PSO, whereby the role of optimisation is to minimise the measures between centroid clusters and data errors. The proposed *opt*-FCMMV will be tested with missing value mechanisms of MAR, MCAR, and MNAR.

A Sample Example

As an example, the infusion of missing values based on MAR, MCAR, and MNAR mechanisms is presented as samples. These examples used the Ovary Cancer dataset. Figure 1 illustrates a sample of 5% missing values of all mechanisms. Meanwhile, Figure 2 shows a sample of 80% missing values of all mechanisms. Both illustrations demonstrated a major level of differences of missing values in the data. The randomly injected missing values for each dataset were calculated based on the percentage of missing values by the total amount of genes in each data according to the missing value mechanisms. The amount of red indicates the number of missing values while green indicates the range of 1. As can be seen in the figures, the illustrated missing values can be expected to be improvised using the proposed method.

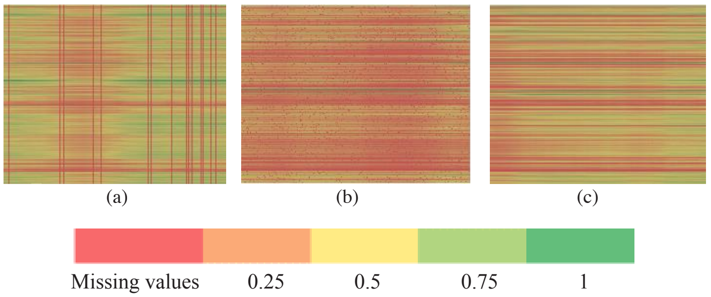

Figure 1. (a) MAR; (b) MCAR; (c) MNAR mechanisms of 5% missing values.

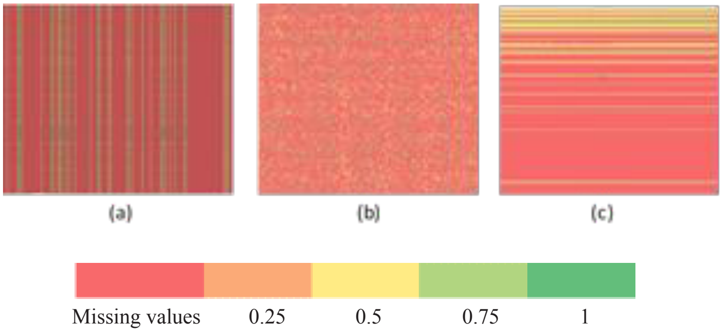

Figure 2. (a) MAR; (b) MCAR; (c) MNAR mechanisms of 80% missing values.

Proposed Method

The central idea of the proposed *opt*-FCMMV method is presented in Figure 3. The figure shows that the algorithm begins with the FCM algorithm hybridised with MV in order to impute values in the gene expression data. Here, MV functions to compare the generated values and aggregate the votes on the values to choose the best values to be imputed. Meanwhile, the imputed values will be initialised with the particles and evaluated. The purpose of the optimisation of PSO is to minimise the error rates and train PSO with a complete dataset in order to estimate the values that correspond to the input of the imputed values by the rule of fitness variance less than threshold values. With this attempt, the best optimised values are imputed in the missing data of the datasets. A detailed explanation is discussed in the upcoming sections.

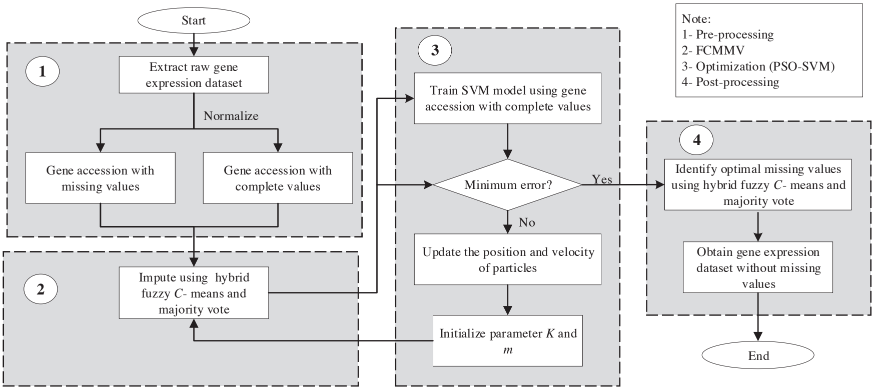
Start
1
Extract raw gene expression dataset
Normalize
Gene accession with missing values
Gene accession with complete values
2
Impute using hybrid fuzzy C- means and majority vote
3
Train SVM model using gene accession with complete values
Minimum error?
Yes
No
Update the position and velocity of particles
Initialize parameter K and m
Note:
1- Pre-processing
2- FCMMV
3- Optimization (PSO-SVM)
4- Post-processing
4
Identify optimal missing values using hybrid fuzzy C- means and majority vote
Obtain gene expression dataset without missing values
End

Figure 3. Proposed *opt*-FCMMV.

Hybrid Fuzzy C-Means with Majority Vote

Based on fuzzy clustering algorithms, an object might belong to more than one cluster with probabilities (Bezdek et al., 1981). The FCM algorithm was originally introduced by Bezdek et al. (1981) and later enhanced by Dunn (1973) to ensure well-separated clusters. However, in this research, FCM will be improvised by hybridising it with MV so that the best imputation values will be identified in the gene expression data. The main steps of the FCMMV imputation method are as follows based on idea of Zhang and Shen (2014).

Step 1: The parameter values of the cluster size and the weighting factor, *m*, are set and the membership function, *U*, is initialised.

Step 2: The cluster centroids are calculated, where $c= \{c_{1,} c_{2,} \;_{...,}\; c_k\}$ based on Equation 1:

$$c_k = \frac{\sum_{i=1}^{n} U[\,X_1(c_k\,)]^m .x_j}{\sum_{i=1}^{n} U[\,X_1(c_k\,)]^m} \tag{1}$$

where c_k = $(1 \le k \le K\,)$ is the *k* th cluster centroid, the parameter, m = $(1 < M\,)$ is the weighting factor (real number) that influences the fuzzy degree of clustering, and the membership function, $U = (x_i,\ c_k)$ is defined as follows in Equation 2 for the cluster centres. For all, $x_{i,}\ \sum_{j=1}^{k} U(\,x_i\,,c_j\,) = 1, i = 1,...,n.$

$$U(\,x_i\,,c_k\,) = \frac{d((\,x_i\,,c_k\,)^{\frac{-2}{m-1}}}{\sum_{j=1}^{k} d((\,x_i\,,c_k\,)^{\frac{-2}{m-1}}} \tag{2}$$

where $d(x_i,c_k)$ is the distance between the data, x_i and the centroid, c_k. This can be calculated through Equation 3:

$$d(\,x_i\,,c_k\,) = (\,\sum_{j=1}^{i} |x_{ij} - c_k\,|^p\,)^{\frac{1}{p}} \tag{3}$$

where p = 2 and p = 1 indicate the Euclidean and Manhattan distances, respectively, and are the cases of Minkowski distances. This research article utilises the value of p = 1.5.

Step 3: The objective function is minimised and defined. The optimal values are searched based on *U* and *C* as stated in Equation 4:

$$J(\,U,C\,) = \sum_{i=1}^{n} \sum_{k=1}^{K} U(\,x_i\,,c_k\,)^m .d(\,x_i\,,c_k\,) \tag{4}$$

Step 4: The termination condition is met if the preset threshold values are more than the objective function values. The difference between the preset thresholds is more than the values of an objective function of two successive iterations or the number of successive iteration reaches the preset threshold's maximum number. Then, the next step is proceeded; otherwise, *U* values have to be updated based on Equation (2) and back to Step 2.

Step 5: The optimal values of *U* and *C* are obtained in order to estimate the missing attribute values of x_i in accordance with Equation 5:

$$\widehat{x}_{ij} = \sum_{k=1}^{K} U(x_i, c_k)(c_k) \tag{5}$$

where $\widehat{x}_{ij}$ represents the missing value that acts as the non-reference attribute.

Step 6: *K* is considered as the target label with $C_i, \forall i \in \Delta = \{1,2,...,K\}$ representing *i* th predicted target label. Given as input *x*, provided with respect to the target labels, yielding a total of *K* predictions, i.e. $P_1,...,P_k$. MV aims to produce a combined predictions of the estimated missing attributes for input *x*, *P*(x)=*j, j* $\in \Delta$from all the *K* predictions, i.e.$P_k\ (x) = j_k, k=1,...,k.$A binary function is used to represent the votes as in Equation 6:

$$V_k(x \in c_i) = \{ \begin{matrix} 1, if \quad p_k(x) = i, i \in \Delta \\ 0, otherwise \end{matrix} \tag{6}$$

The sum of the votes from all *K* for each C_i and the label that receives the highest *gbest* vote are the final phase of estimating missing values of the predicted class. If failed to get the highest vote, then, return to Step 4 till the highest vote is obtained to select the best missing values in the data.

Optimised Hybrid of Fuzzy C-Means with Majority Vote

For the optimisation, Particle Swarm Optimisation and Support Vector Machine (PSOSVM) is selected for this research due to the strong optimisation bond between both methods based on Salleh and Samat's (2017) work on the PSO algorithm. Three steps are used on each gene attribute one by one and the attribute outputs are combined into the output that corresponds to the input. Therefore, the SVM model is trained, "input gene attribute values = output gene attribute values". *opt*-FCMMV is the missing value imputation method proposed in this article. The imputation of FCMMV is to identify the missing values in the dataset, whereby the parameters *K* and *m* are optimised (with the assistance of PSOSVM) with the best *K* votes. The purpose of the PSO algorithm with SVM in this research is to minimise the error rate. The objective function is minimised via *(Input-Output)* ^{2}, where the input is the FCMMV imputation and the output is the SVM prediction. Before the final optimal imputation of the missing values in the dataset, SVM must be trained with a complete dataset in order to recall and estimate the values that correspond to the input.

Step 1: The datasets without any missing values are the samples that will be selected.

Step 2: One of the input gene attributes are set, some of the values that are missing act as the output gene attributes, which are also the condition gene attributes.

Step 3: SVM is used to predict each value of gene attribute.

Step 4: X_c represents the complete data, while X_m represents the missing data. The input is as shown in Equation 7 and the output is as shown in Equation 8:

$$Input = \begin{pmatrix} X_c \\ X_m \end{pmatrix} \tag{7}$$

$$Output = f\begin{pmatrix} X_c \\ X_m \end{pmatrix} \tag{8}$$

where f represents the mapping between the input and output of the SVM model.

Step 5: The input data are recalled in the SVM model and the difference is known as the error. PSO is used to minimise the error between the input and output of the SVM model as shown in Equation 9. The objective function has the responsibility to minimise the error that results in an approximate value for the missing value. Following Equation 10, it shows the objective function of PSO and the outputs are used to minimise the objective function values for completeness.

$$Error = Input - Output \tag{9}$$

$$PSO\ objective\ function = (Input - Output)^2 \tag{10}$$

EXPERIMENTAL RESULTS

This study empirically evaluated *opt*-FCMMV by comparing its performance with FCM and FCMMV algorithms. Experiments were conducted on a total of fifteen datasets in the biomedical domain. In Experiment 1 and Experiment 2, a comparison was made between *opt*-FCM with FCM and FCMMV using SVM classifier based on different levels of missing values to examine the efficiency of the proposed method. The SVM classifier was used based on the default parameter values using the Radial Basis Kernel (RBF) (Wahyudi et al., 2010) provided in the LibSVM software package. In Experiment 1, the research discussed on the Ovary Cancer dataset, whereas Experiment 2 elaborated on the Lung Cancer dataset. The differences between the methods MAR, MCAR, and MNAR were calculated based on accuracy rates and Root Mean Squared Error (RMSE). The formulae used are as in Equations 11 and 12 (Shcherbakov et al., 2013; Kouchaki et al., 2018):

$$\text{Accuracy} = (TP + TN)/\ (TP + TN + FP + FN) \tag{11}$$

where TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ y_i - \bar{y}_i\ )^2} \tag{12}$$

where y_i is the original value, is the mean of observed data, and n is the total amount of predictions.

Datasets

To verify the efficiency and effectiveness of *opt*-FCMMV, Experiment 1 consisted of a total of 15 sub-datasets, created from the Ovary Cancer dataset (Zhu et al., 2007). The dataset contained 15,154 genes and 254 instances with two classes, Normal and Cancerous. The dataset used for this research was a normalised dataset without any missing values. The statistics of the dataset is summarised in Table 1. The sample of the dataset is illustrated in Table 2.

Table 1

Description of Ovary Cancer Dataset

Classes	Acronyms	Total instances	Total genes
Normal	N	91	15,154
Cancerous	C	162	

For the optimisation, PSOSVM was selected for this research due to the strong optimisation bond between both methods. Three steps were used on each gene attribute one by one and the attribute outputs were combined into the output that corresponded to the input. Therefore, the SVM model was trained, “input gene attribute values = output gene attribute values”. *opt*-FCMMV is the novel missing value imputation method proposed in this article. The imputation of FCMMV is to identify the missing values in the dataset, whereby the parameters *K* and *m* are optimised (with the assistance of PSOSVM) with the best *K* votes. On the other hand, the purpose of the PSO algorithm with SVM is to minimise the error rate. The objective function was minimised via *(Input-Output)* ^{2}, where the input is the FCMMV imputation and the output is the SVM prediction. Before the final optimal imputation of the missing values in the dataset, SVM must be trained with a complete dataset in order to recall and estimate the values that corresponded to the input.

Table 2

Ovary Cancer Data Sample

MZ-7.86E-05	MZ2.18E-07	…	MZ19995.513	Class
0.494626	0.263735	…	0.449296	N
0.258063	0.406593	…	0.619718	N
0.537636	0.032966	…	0.035918	N
0	0.395605	…	0.486621	N
0.526884	0.395605	…	0.251408	N
0.39785	0.395605	…	0.333102	N
0.64516	0.307689	…	0.567607	N
0.720432	0.351644	…	0.46268	N
0.537636	0.307689	…	0.567607	N
0.526884	0.494503	…	0.350704	N
0.720432	0.351644	…	0.564088	N
…	…	…	…	…
0.763442	0.527469	…	0.464088	C
0.569893	0.681316	…	0.498592	C
0.569893	0.791209	…	0.450005	C
0.688175	0.703294	…	0.519718	C
0.838709	0.824175	…	0.519718	C
0.795699	0.64835	…	0.273243	C

Experiment 2 consisted of a total of 15 sub-datasets that were created from the Lung Cancer dataset (Zhu et al., 2007) with five classes. The dataset contained 12,600 genes and 204 instances. The dataset used for this research was a complete and non-normalised dataset without any missing values. The statistics of the dataset is summarised in Table 3.

Table 3

Description of Lung Cancer Dataset

Classes	Acronyms	Total instances	Total genes
Normal	N	17	12,600
Lung Adenocarcinomas	LC	139	
Pulmonary Carcinoids	PC	20	
Small Cell Lung Carcinomas	SCLC	6	
Squamous Cell Lung Carcinomas	SQCLC	21	

The sample of the dataset is illustrated in Table 4. The dataset was normalised using the following Equation 14 (Wenzel & Peter, 2017) within the range of [0, 1] to reduce redundancies and data anomalies.

$$x_n = \frac{x_o - x_{min}}{{}^{x}max - {}^{x}min} \tag{14}$$

where x_n = the new value for variable X, x_o = the current value for variable X, x_{min} = the minimum data point, and x_{max} = the maximum data point in the dataset .

Table 4

Lung Cancer Data Sample

AFFX-MurIL2_at	AFFX-MurIL10_at	…	109_at	Class
-18.6	10.54	…	76.98	LA
9.12	9.12	…	105.73	LA
-2.175	-2.21	…	73.735	LA
-1.54	21.75	…	65.435	LA
-9.07	3.08	…	39.54	LA
-16.58	-20.09	…	59.49	LA
-15.895	10.88	…	39.965	LA
-14.5	-10.48	…	96.35	LA
-25.595	2.175	…	60.53	LA
1.23	24.74	…	52.95	LA
-13.95	12.41	…	51.62	LA
…	…	…	…	…

(continued)

AFFX-MurIL2_at	AFFX-MurIL10_at	…	109_at	Class
1.68	-7.45	…	82.21	PC
35.14	106.16	…	118.41	PC
-21.15	-31.2	…	65.03	PC
26.9	10.44	…	71.97	PC
23.8	29.14	…	135.08	PC
-18.37	-1.03	…	40.17	PC

Experiment 1

This section reports Experiment 1 with the Ovary Cancer dataset conducted for the efficiency of the *opt*-FCMMV performance through the comparisons before the enhancement of *opt*-FCMMV, such as with no imputations, FCM, and FCMMV. Results of Experiment 1 are included in Table 5, which consist of accuracy rates based on the methods used on missing values in the data. Figures 3 to 5 show the RMSE rates with their missing ratios. Accuracy rates (Acc.) are used to identify whether the role of the methods is worthwhile to handle the missing values, while RMSE covers for prediction errors. For the MAR mechanism, there were 758 to 12,123 missing values in the data (refer Table 5). It can be seen that the proposed method showed the highest accuracy rates with 93.7% for 5% and 10% missing values, 96.8% for 30% missing values, 95.3% for 50% missing values, and 98.0% accuracy rates for 80% missing values comparable to other methods.

Table 5 also depicts the MCAR scenarios. The proposed method showcased higher accuracy rates as well. There was a hike from 85.4% to 87.0% using the *opt*-FCMMV method for 5% missing values (refer Table 5). The same goes to all other rates of missing values with high accuracy rates. Additionally, for the MNAR scenario, there were missing data columns from 5% (13) to 80% (203). Furthermore, the accuracy rates improvised from 83.0% using FCM with missing values to 94.1% using the proposed method (refer Table 5). Among the different experiments from Table 5, this indicated that the best performance of the proposed method with highest accuracy rates was shown using SVM. With *opt*-FCMMMV, all the experiments with different methods and missing value rates (5%, 10%, 30%, 50%, and 80%) demonstrated significant performance improvement as compared to other methods before enhancement (FCM and FCMMV). Here, MV assisted to consider the most suitable values to be imputed for the highest voted values. Optimisation was able to assist the imputed values into many feasible values and the best-predicted values were able to be selected for the missing regions for the respective mechanisms.

Table 5

Accuracy Rates of Data with Different Missing Value Mechanisms

MAR mechanism					
MV	Missing values	None	FCM	FCMMV	*opt*-FCMMV
5%	758	64.00	81.80	90.90	**93.70**
10%	1,515	64.00	90.70	91.30	**93.70**
30%	4,546	36.00	80.20	94.50	**96.80**
50%	7,577	36.00	65.50	94.90	**95.30**
80%	12,123	36.00	64.30	65.50	**98.00**
MCAR mechanism					
5%	192,456	85.40	70.20	85.80	**87.00**
10%	384,912	79.80	72.70	81.00	**84.20**
30%	1,154,735	64.80	66.80	71.00	**72.70**
50%	1,924,558	69.20	71.10	77.10	**93.30**
80%	3,079,293	64.00	72.70	91.30	**91.80**
MNAR mechanism					
5%	13	89.70	83.00	86.60	**94.10**
10%	25	86.60	83.40	86.60	**88.10**
30%	76	86.20	81.40	81.80	**90.50**
50%	127	84.20	79.10	84.60	**86.60**
80%	203	75.10	75.10	76.70	**79.40**

Figures 4 to 6 show the RMSE rates for all missing value mechanisms. The lower rate of RMSE indicated that it was in a better fit. Most Ovary Cancer data with missing value results were in the lower RMSE rates, which successfully approved the proposed *opt*-FCMMV method. With the perspective of mechanisms for the Ovary Cancer data, it can be seen that *opt*-FCMMV as the proposed method was able to show high accuracy differences with other methods for the MAR mechanism as compared to the MCAR and MNAR mechanisms.

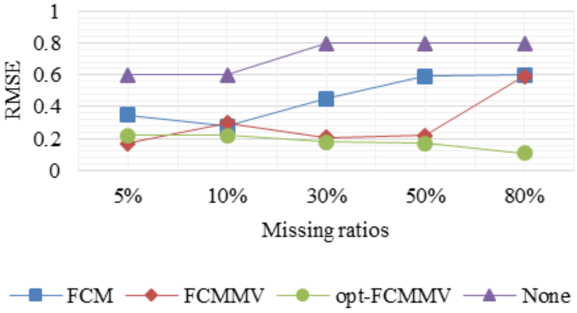

Figure 4. Ovary Cancer data: RMSE rates based on missing ratios of MAR mechanism.

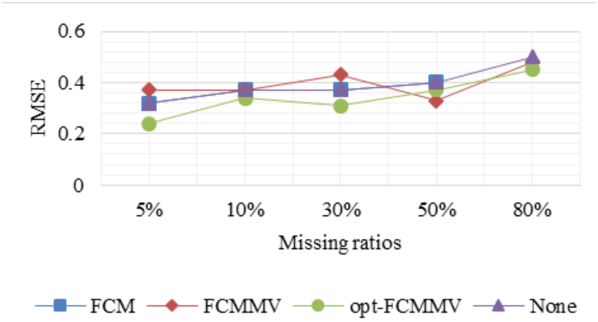

Figure 5. Ovary Cancer data: RMSE rates based on missing ratios of MCAR mechanism.

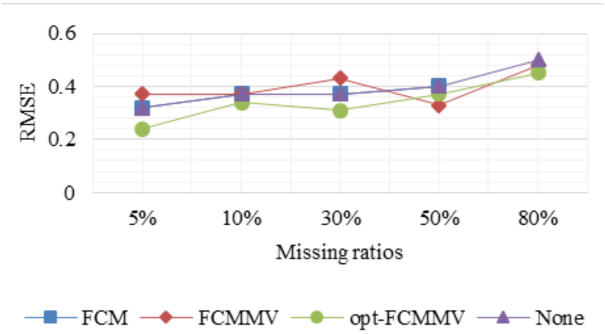

Figure 6. Ovary Cancer data: RMSE rates based on missing ratios of MNAR mechanism.

Experiment 2

This section reports Experiment 2 with the Lung Cancer dataset to identify the effectiveness of the proposed method. Tables 8 till 10 for Experiment 2 illustrate different missing value ratios and comparisons between methods such as no methods used to handle missing values, FCM, FCMMV, and *opt*-FCMMV. As shown in Table 6, with the increase in missing ratios, *opt*-FCMMV was able to demonstrate promising and high accuracy rates. For 80% of missing ratio, the proposed method was able to improvise 20.2% from the initial accuracy with no imputation and FCM methods. Referring to Table 6, *opt*-FCMMV showed higher accuracy rates as compared to other methods. 5% missing ratio results indicated that FCMMV obtained 72.9%, which was higher as compared to *opt*-FCMMV's 71.4% accuracy rate. This is due to the poor measure of the MNAR mechanism via 5% of missing value ratio.

Table 6

Accuracy rates of data with different mechanisms of Lung Cancer data

MAR mechanism					
MV	Missing values	None	FCM	FCMMV	*opt*-FCMMV
5%	630	68.5	71.40	72.40	**73.20**
10%	1,260	68.5	72.40	75.40	**75.90**
30%	3,780	68.5	79.80	81.00	**84.20**
50%	6,300	68.5	68.50	85.20	**87.20**
80%	10,080	68.5	68.50	87.70	**88.70**
MCAR mechanism					
5%	127,890	68.5	68.50	70.90	**72.90**
10%	255,780	68.5	68.50	73.90	**74.40**
30%	767,340	68.5	68.50	71.40	**74.40**
50%	127,890	68.5	69.80	70.20	**75.10**
80%	2,046,240	68.5	70.00	70.20	**72.90**
MNAR mechanism					
5%	10	68.5	68.50	**72.90**	71.40
10%	20	68.5	71.90	74.40	**85.70**
30%	61	68.5	61.60	79.80	**84.20**
50%	102	68.5	69.00	75.40	**77.30**
80%	162	68.5	56.20	58.10	**68.50**

Figures 7 to 9 show that *opt*-FCMMV had lower RMSE values, indicating the significance of the proposed method as compared to other methods. Lower values of RMSE by *opt*-FCMMV could be due to the presence of a small number of high error predictions, which showcased the efficiency of the proposed method, *opt*-FCMMV. It can be deduced that the MAR mechanism showed a good comparison of accuracy rates and stable experimental results as compared to other mechanisms.

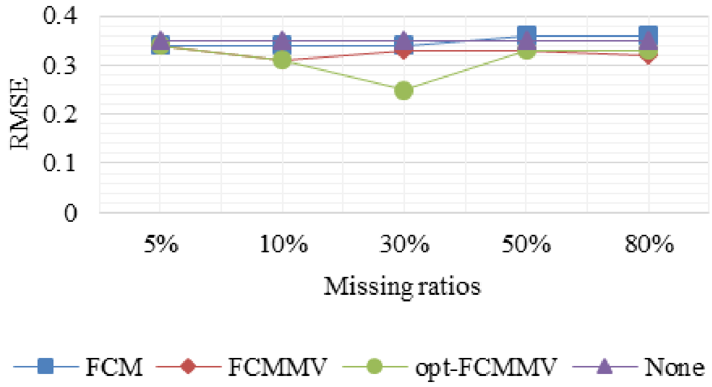

Figure 7. Lung Cancer data: RMSE rates based on missing ratios of MAR mechanism.

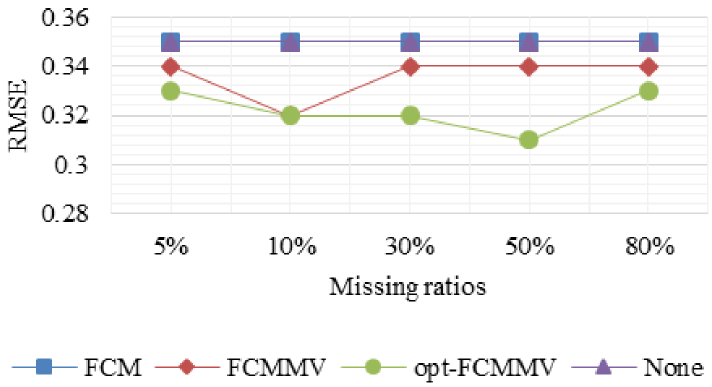

Figure 8. Lung Cancer data: RMSE rates based on missing ratios of MCAR mechanism.

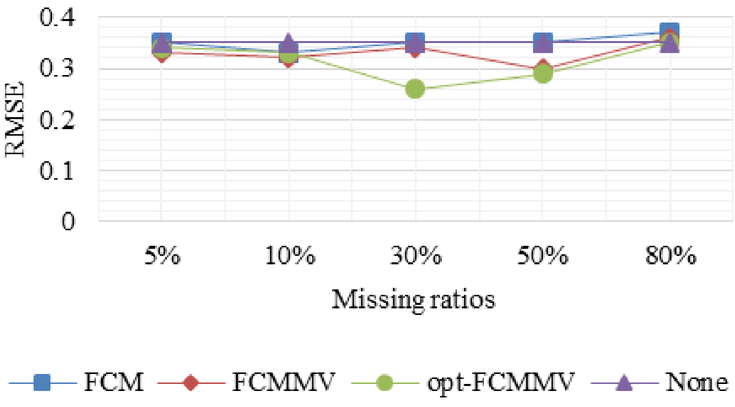

Figure 9. Lung Cancer data: RMSE rates based on missing ratios of MNAR mechanism.

DISCUSSION

In this article, both experiments used two existing methods, which are no imputations and FCM. Mewnwhile, the improved FCM such as FCMMV and *opt*-FCMMV were also utilised to evaluate the accuracy rates of the imputation method on microarray data. Accuracy rate and RMSE were used to measure the credibility of the algorithm since RMSE can show the increase and decrease in methods by the increase in sample size with any missing value rate. Accuracy rate was used to measure the performance of the methods because the quantity of information missed increased due to the number of missing values, whereby it led to affecting the accuracy rates.

For Experiment 1, all accuracy rate results from the experiments showed that *opt*-FCMMV was the best method to impute the missing values. In Experiment 2, 14 out 15 experiments proved the credibility of the proposed method based on accuracy rates. These results from the experiments showed the ability of the proposed method in imputing the missing values in the data whether in smaller or larger ratio. While for RMSE values, almost all mechanisms (MCAR, MNAR, and MAR) for Experiments 1 and 2 showed the lowest values, proving the methods' advantages. One of the major advantages of the proposed method is that the algorithm used the information from the data itself to predict the missing values. This is also due to MV that assisted in choosing the best optimal measurement for gene similarity. Another advantage in this method is the optimisation itself. Optimising the coefficients of the non-missing values of the similar genes via the proposed method allowed to gain the nearest gene measurements in accordance with the class of the genes. Furthermore, this method worked well for a large number of missing values. This is due to the PSO algorithm's search strategy as it minimised the error rate that directly improved the accuracy rates and lowered the RMSE values.

CONCLUSION

In this article, a new imputation method, known as Optimised Hybrid of Fuzzy C-Means and Majority Vote (*opt*-FCMMV) was proposed. This new method created a more solid and informative dataset as compared to other methods due to its optimisation method. Therefore, the achieved accuracy rates are higher through the improved method from FCM, FCMMV to *opt*-FCMMV. The experimental results confirmed the proposed method can be a credible method for upcoming research in handling missing values. In this article, the proposed method was compared against three imputation methods (i.e. None, FCM, and FCMMV), with five types of missing value percentage (i.e. 5%, 10%, 30%, 50%, and 80%). The Ovary and Lung Cancer microarray data were used as datasets that covered the biomedical field. The *opt*-FCMMV method has proven that it can solve high dimensional problems and improve accuracy across different types of missing value percentage. In the future, *opt*-FCMMV can also be applied in different domains while other imputation methods and metaheuristic algorithms for optimisation can be investigated. *opt*-FCMMV can be considered as a promising imputation method for the pre-processing stage for future research in the biomedical field.

ACKNOWLEDGEMENT

The authors would like to acknowledge the financial support from the School of Graduate Studies, Universiti Teknologi Malaysia (Zamalah Scholarship) and Research University Grant, Tier 1 under the Vote No. 16H69.

REFERENCES

Baraldi, P., Di Maio, F., Genini, D., & Zio, E. (2015). Reconstruction of missing data in multidimensional time series by fuzzy similarity. *Applied Soft Computing*, *26*, 1–9. https: //doi.org/ 10.1016/j.asoc.2014.09.038

Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2017). From predictive methods to missing data imputation: An optimization method. *The Journal of Machine Learning Research*, *18*(1), 7133–7171. Retrieved from http:// jmlr.org/papers/v18/17-073.html

Bezdek, J. C., Coray, C., Gunderson, R., & Watson, J. (1981). Detection and characterization of cluster substructure I. Linear structure: Fuzzy c-lines. *Siam Journal on Applied Mathematics*, *40*(2), 339–357.

Bose, S., Das, C., Gangopadhyay, T., & Chattopadhyay, S. (2013, December). A modified local least square based missing value estimation method in microarray gene expression data. In *2013 2nd International Conference on Advanced Computing, Networking and Security*, Mangalore, India (pp. 18–23). https: //doi.org/ 10.1109/ADCONS.2013.11

Chattopadhyay, S., Das, C., & Bose, S. (2015, December). A novel biclustering based missing value prediction method for microarray gene expression data. In *2015 International Conference on Man and Machine Interfacing (MAMI)*, Bhubaneswar, India (pp. 1–6). https: //doi.org/ 10.1109/MAMI.2015.7456603

De Silva, H. M., & Perera, A. S. (2017). Evolutionary k-nearest neighbor imputation algorithm for gene expression data. *International Journal on Advances in ICT for Emerging Regions (ICTER)*, *10*(1), 1–8. https: // doi.org/ 0.4038/icter.v10i1.7183

Dibal, N. P., Okafor, R., & Dallah, H. (2017). Challenges and implications of missing data on the validity of inferences and options for choosing the right strategy in handling them. *International Journal of Statistical Distributions and Applications*, *3*(4), 87–94. https: //doi.org/ 10.11648/j.ijsd.20170304.15

Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 32–57. https: //doi.org/ 10.1080/01969727308546046

Gupta, A., Wang, H., & Ganapathiraju, M. (2015, November). Learning structure in gene expression data using deep architectures, with an application to gene clustering. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Washington, DC, USA (pp. 1328–1335). https: //doi.org/ 10.1109/BIBM.2015.7359871

Hourani, M. A., & El Emary, I. M. (2009). Microarray missing values imputation methods: Critical analysis review. *Computer Science and Information Systems*, *6*(2), 165–190. https: //doi.org/ 10.2298/csis0902165H

Keerin, P., Kurutach, W., & Boongoen, T. (2012, October). Cluster-based kNN missing value imputation for DNA microarray data. In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC*), Seoul, South Korea (pp. 445–450). https://doi.org/ 10.1109/ICSMC.2012.6377764

Keerin, P., Kurutach, W., & Boongoen, T. (2016). A cluster-directed framework for neighbour based imputation of missing value in microarray data. *International Journal of Data Mining and Bioinformatics*, *15*(2), 165–193. https: //doi.org/ 10.1504/IJDMB.2016.076535

Kellermann, A. P., Trevathan, D., & Kromrey, J. D. (2016). Missing data and complex sample surveys using SAS®: The impact of listwise deletion vs. multiple imputation on point and interval estimates when data are MCAR and MAR, 1–12. Retrieved from https://pdfs.semanticscholar.org/2018/9420dd222e9ba823aadfe56bc4fdf00d3f58.pdf?_ga=2.232370207.1411031538.1586998037-1299407029.1586998037

Kouchaki, S., Yang, Y., Walker, T. M., Walker, A. S., Wilson, D. J., Peto, T. E., & Cryptic Consortium. (2018). Application of machine learning techniques to tuberculosis drug resistance analysis. *Bioinformatics*, 1–7. https: //doi.org/ 10.1093/bioinformatics/bty949

Lazar, C., Gatto, L., Ferro, M., Bruley, C., & Burger, T. (2016). Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies. *Journal of Proteome Research*, *15*(4), 1116–1125. https: //doi.org/ 10.1021/acs.jproteome.5b00981

Li, Y., Ngom, A., & Rueda, L. (2010, May). Missing value imputation methods for gene-sample-time microarray data analysis. In *2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, Montreal, QC, Canada (pp. 1–7). https: //doi.org/: 10.1109/CIBCB.2010.5510349

Marwala, T. (2009). Optimization methods for estimation of missing data. In *Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques* (pp. 210–232). https: //doi.org/ 10.4018/978-1-60566-336-4.ch010

Ouyang, M., Welsh, W. J., & Georgopoulos, P. (2004). Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, *20*(6), 917–923. https: //doi.org/10.1093/bioinformatics/bth007

Paul, A., Sil, J., & Mukhopadhyay, C. D. (2017). Gene selection for designing optimal fuzzy rule base classifier by estimating missing value. *Applied Soft Computing*, *55*, 276–288. https: //doi.org/ 10.1016/j.asoc.2017.01.046

Pino Angulo, A. (2018). Gene selection for microarray cancer data classification by a novel rule-based algorithm. *Information*, *9*(1), 1–15. https: //doi.org/ 10.3390/info9010006

Pourhasem, M. M., Kelarestaghi M., & Pedram, M. M. (2010) Missing values estimation in microarray data using fuzzy clustering and semantic similarity. *Global Journal of Computer Science and Technology*, *12*(10), 18–22.

Qin, F., & Lee, J. (2010, December). Dynamic methods for missing value estimation for DNA sequences. In *Computational and Information Sciences*, Chengdu, China (pp. 442–445). https: //doi.org/10.1109/ICCIS.2010.115

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https: //doi.org/10.1093/biomet/63.3.581

Salleh, M. N. M., & Samat, N. A. (2017, August). FCMPSO: An imputation for missing data features in heart disease classification. In *IOP Conference Series: Materials Science and Engineering*, 226(1), Melaka, Malaysia (p. 012102). https: //doi.org/ 10.1088/1757-899X/226/1/012102

Saha, S., Ghosh, A., Seal, D. B., & Dey, K. N. (2016). An improved fuzzy based missing value estimation in DNA microarray validated by gene ranking. *Advances in Fuzzy Systems*, 1–19. https: //doi.org/ 10.1155/2016/6134736

Shcherbakov, M. V., Brebels, A., Shcherbakova, N. L., Tyukov, A. P., Janovsky, T. A., & Kamaev, V. A. E. (2013). A survey of forecast error measures. *World Applied Sciences Journal*, *24*(24), 171–176. https: //doi.org/ 10.5829/idosi.wasj.2013.24.itmies.80032

Shehab, M., Khader, A. T., & Laouchedi, M. (2018). A hybrid method based on cuckoo search algorithm for global optimization problems. *Journal of Information and Communication Technology*, *17*(3), 469–491.

Suphanchaimat, R., Limwattananon, S., & Putthasri, W. (2017). Multiple imputation technique: Handling missing data in real world health care research. *Southeast Asian Journal of Tropical Medicine and Public Health*, *48*(3), 694–703. Retrieved from http://www.tm.mahidol.ac.th/seameo/journal-48-3-2017.html

Suyundikov, A., Stevens, J. R., Corcoran, C., Herrick, J., Wolff, R. K., & Slattery, M. L. (2015). Accounting for dependence induced by weighted kNN imputation in paired samples, motivated by a colorectal cancer study. *PloS One*, *10*(4), 1–15. https: //doi.org/10.1371/journal.pone.0119876

Tian, J., Yu, B., Yu, D., & Ma, S. (2012). A fuzzy clustering method for missing value imputation with non-parameter outlier test. *Information Science and Industrial Applications*, 33–42.

Tsai, C. F., Li, M. L., & Lin, W. C. (2018). A class center based approach for missing value imputation. *Knowledge-Based Systems*, *151*, 124–135. https: //doi.org/ 10.1016/j.knosys.2018.03.026

Tshering, S., Okazaki, T., & Endo, S. (2013). A method to identify missing data mechanism in incomplete dataset. *International Journal of Computer Science and Network Securit*y, *13*(3), 14–22. Retrieved from http://paper.ijcsns.org/07_book/201303/20130303.pdf

Tung, A. K., Zhang, R., Koudas, N., & Ooi, B. C. (2006, September). Similarity search: A matching based method. In *Proceedings of the 32nd International Conference on Very Large Data Bases*, Seoul, South Korea (pp. 631–642). https: //doi.org/ 10.1.1.111.3244

Wahyudi, G., Fanany, M. I., Jatmiko, W., & Arymurthy, A. M. (2010, November). SVM kernels accuracy and generalization capability on apnea detection from ECG. In *Proceedings of the 2nd International Conference on Advanced Computer Science and Information Systems*, Bali, Indonesia (pp. 187–191). Retrieved from https://pdfs.semanticscholar.org/82b1/7a4790a90f4022b8f7c4bfa29bae61bb77c2.pdf

Wenzel, S., & Peter, T. (2017). Comparing different distance metrics for calculating distances in urban areas with a supply chain simulation tool. *Simulation in Produktion und Logistik 2017*, 119. https: //doi.org/ 10.19211 / KUP9783737601931

Yaraghi, S., Jazi., M.D., & Rafeh, V. (2012). Missing value estimation in microarray data by fuzzy clustering and gene regulatory information. In *4th International Conference on Bioinformatics and Biomedical Technology*, Singapore, 29 (pp. 19–23). Retrieved from http://www.ipcbee.com/vol29/4-ICBBT2012-H011.pdf

Yu Z., Li T., Horng S.J., Pan. Y., Wang, H., & Jing, Y. (2017). An iterative locally auto-weighted least squares method for microarray missing value estimation. *IEE Transactions on Nanobiosciences*, *16*(1), 21–33. https: //doi.org/ 10.1109/TNB.2016.2636243

Zhang, J., & Shen, L. (2014). An improved fuzzy C-means clustering algorithm based on shadowed sets and PSO. *Computational Intelligence and Neuroscience*, *2014*, 1–11. https: //doi.org/ 10.1155/2014/368628

Zhu, Z., Ong, Y. S., & Dash, M. (2007). Markov blanket-embedded genetic algorithm for gene selection. *Pattern Recognition*, *40*(11), 3236–3248. https: //doi.org/ 10.1016/J.Patcog.2007.02.007.