Research Article

Missing data imputation with fuzzy feature selection for diabetes dataset

Mohamad Faiz Dzulkalnine[1] · **Roselina Sallehuddin**[1]

Abstract
Missing data in datasets remain as a difficulty in terms of data analysis in various research fields, especially in the medical field, as it affects the treatment and diagnosis that the patient should receive. In this research, Fuzzy c-means (FCM) are used to impute the missing data. However, like in most data imputation methods, FCM do not consider the presence of irrelevant features. Irrelevant features can increase the computational time of the imputation process and decrease the accuracy of the prediction. Feature selection techniques can alleviate this problem by selecting the most relevant features and reducing the dataset size. Fuzzy principal component analysis (FPCA) is used as the feature selection method in this study as it considers the presence of outliers compared to classical PCA as outliers are the main reason some features renders irrelevant. Therefore, an improved hybrid imputation model of FPCA–Support vector machines–FCM (FPCA–SVM–FCM) has been proposed and employed in this study. The efficiency of the proposed model is investigated on one dataset which is Pima Indians Diabetes dataset. Experimental results showed that the proposed hybrid imputation model is better than the existing methods by producing a more accurate estimation in terms of accuracy, RMSE and MAE. The proposed method was also validated by using Wilcoxon rank sum and Theil's *U* test and obtained good results compared to SVM–FCM. Therefore, it can be used as an alternative tool for handling missing data in order to obtain a better quality dataset.

Keywords Missing data · Fuzzy feature selection · Imputation · Classification

1 Introduction

Missing data are unwanted in machine learning and data mining as missing data pose many problems. Missing data occur in datasets for several reasons, for example, malfunctions of equipment, non-response in surveys, insufficient resolution, image corruption, incorrect measurements, dust or scratches on the slides, incorrect entering of data, or experimental error in the laboratory procedure. Missing data can be categorized into three types, which are missing completely at random (MCAR), missing not at random (MNAR) and missing at random (MAR) [1]. MCAR is where the missing data have no relationship with the variable. It means that the missing data do not depend on any other variable. The second type of missing data is MNAR where the missing data have a relationship with the other missing data. In the case of MNAR, the missing data cannot be estimated from existing variables. The third and last type of missing data is MAR where the missing data has a relationship with other variables. If data in a dataset is MAR, the missing data can be predicted by using other variables. This means there is a probability that the missing data are dependent on the value of other variables [2]. In this study, we assumed that data are MAR which implies the missing data can be predicted by utilizing information from the remaining data.

Missing data raised some issues in data analysis which are loss of precision due to fewer data available, and bias

✉ Mohamad Faiz Dzulkalnine, m.faiz.dzul@gmail.com; Roselina Sallehuddin, roselina@utm.my | [1]Faculty of Computing, Universiti Teknologi Malaysia, 81300 Skudai, Johor, Malaysia.

SN Applied Sciences (2019) 1:362 | https://doi.org/10.1007/s42452-019-0383-x

Received: 1 January 2019 / Accepted: 19 March 2019 / Published online: 26 March 2019

due to distortion of the data distribution. Some decision-making tools such as ANN, SOM, SVM, and other computational interface techniques cannot be employed if data are not complete. Missing data in the medical dataset, on the other hand, raise the issue in the process of creating conclusion from case files. Missing data pose much greater concern, especially when the conclusion will affect the correct attention a patient should receive. For example, in cancer prognosis, it is important to discover the cancer relapse of a particular patient and the decision-making process involved in the patient's treatment. Missing data could reduce the number of available cases for analysis or even distort the analysis that is caused by biases during the estimation process.

Accuracy is a major issue in handling missing data in the dataset as it can affect the reliability of the data analysis results [3]. The accuracy of diagnosis of patient's disease such as diabetes, breast cancer, and others is greatly depending on experts' experience. Nevertheless, missing data present in the patient's data can diverge the decision made from the experts [4]. Moreover, missing data will create bias that leads to the misleading results [5]. Previous methods such as deleting, ignoring, zero or mean estimation are likely to introduce bias, especially when the missing rates are high [6]. Traditional imputation method such as deletion will introduce bias because the subsample of attributes represented by the missing data is not representative of the original sample.

Previous popular imputation methods in handling missing data include KNN, BPCA, and SVDimpute. The SVDimpute and BPCA which are global imputation methods perform better only when the datasets are homogenous. It means that the imputation estimation will be less accurate when there are dominant local similarity structures among the data [7]. The imputation performance of KNN, on the other hand, a local imputation method, will severely get affected if the data are globally correlated instead of locally correlated.

Consequently, a new hybrid imputation algorithm that consists of fuzzy c-means with support vector regression and a genetic algorithm were proposed [8] to handle missing data in datasets. Aydilek et al. [8] introduce training phase in their proposed imputation model so that it can obtain the difference of error between the imputed dataset with the trained dataset. The hybrid algorithm shows excellent result of imputation by incorporating training before the imputation process. However, in the training process, the presence of irrelevant features or data can affect the accuracy of imputation and increases the bias. The author suggested a feature selection method before the training phase to increase the imputation accuracy.

Aydilek et al.'s [8] suggestion also has been supported by some previous study. In 2008, a study has shown that uncorrelated features will reduce the imputation efficiency as previous imputation method such as traditional KNN tends to bias toward outliers or uncorrelated features. As a consequence, the performance of KNN degrades, especially when the missing rate increases [9]. The paper proposes feature selection before imputation which is the modified KNN called KNN-based feature selection (KNN-FS) [10]. It is then found out that, by implementing feature selection before imputation, the proposed method performed better than traditional KNN in terms of NRMSE when applied to three microarray datasets: Lung Tumor, Colon Cancer, and ALL-AML Leukemia dataset.

Another study published in 2012 [11] also applied feature selection before imputation. The authors compared the performance of mutual information estimators with or without feature selection before the imputation step. The experiment result shows that by selecting significant features before imputing the data generally increases the accuracy of the prediction models, especially when the missing rate is high. Their approach indicates that by using real-world datasets such as Delve, Nitrogen, and Housing and Mortgage dataset, imputing missing data after the feature selection step will produce accurate prediction models. Both of these studies showed the importance of feature selection before the imputation process. Therefore, in this study, a feature selection method will be employed before the imputation step. Classification is used to measure the accuracy of the selected features in order to determine the relevancy of each feature before the imputation process.

One prominent example of a method that is used frequently for data analysis and pre-processing is principle component analysis (PCA) [12]. Principal components analysis (PCA) is known to be able to select relevant features by removing irrelevant features. PCA was shown to be able to select relevant features from a set of simulated auxiliary variables by reducing the number of auxiliary variables without increasing bias [13]. The inclusion of too many additional features may also introduce bias and decrease the precision due to overfitting. This happens when the features' outcome correlation and sample size are low [14]. By selecting a set of relevant features, the complexity of the process can be reduced and the performance of the learning methods can be improved [15].

Hence, PCA was employed as a feature selection method in credit scoring data and was proved to be a better feature selection method in comparison with other methods such as genetic algorithm (GA), information gain ratio, and relief attribute evaluation function [16]. The study also showed that hybrid of PCA with SVM produces greater classification accuracy when compared to hybridization of PCA with ANN, Naïve Bayes, and Decision Tree. This proves that the combination between PCA

and SVM can produce a good classification method. As this study focuses on medical data, SVM has been shown to perform very well in classifying various types of cancer data [17–19]. With the above-mentioned advantages, SVM is used as the classifier in this study.

Although hybrid of PCA with SVM can potentially produce greater classification accuracy, PCA has one major weakness, which is sensitivity to outliers. This drawback can affect the accuracy performance of feature selection in classification. The sensitivity to outliers can be diminished by incorporating fuzzy element in the calculation of the covariance matrix of the PCA. Fuzzy membership is known to deal with the issue of outliers, and this has been proven in some studies that have applied fuzzy methods in regression analysis. Improvement is due to the reason that feature space is divided differently as a result of nonlinearity in comparison with linear fuzzy PCA. Fuzzy PCA can also reduce the training time as it is proved that fuzzy PCA is a lot faster than classical PCA [20].

In this study, we proposed an imputation method by FCM with feature selection by fuzzy PCA and SVM. The rest of this paper is organized in the following manner. Section 2 will cover on the literature review and related works. Section 3 explains the implementation of the proposed model. Section 4 discusses the experimental data and presents the result of the experiment. Finally, Sect. 5 provides the summary and conclusion.

2 Related methods

In this section, discussions on the FPCA's literature and the implementation FPCA for ranking the relevant features based on weight are described. Next, Support Vector Machines that used as classifier are explained. Finally, the imputation method that is used in this study which is fuzzy c-means is also discussed.

A. Fuzzy principal component analysis

Classical PCA calculates the eigenvalues, and its corresponding eigenvectors are determined by the covariance matrix. To obtain the covariance matrix, PCA must first calculate the mean of the attributes. However, PCA has two weaknesses. Firstly, classical PCA calculates the eigenvector in "batch way" but real-world application is usually analyzed incrementally or "on-line" way. This proved problematic when a new sample is added and PCA is unable to adapt to the new data. The second weakness is that PCA only performed well on datasets that do not contain any outlier, yet real-world dataset realistically will contain outliers and it is still a challenging task for researchers to separate the outliers from the dataset. Outliers can be a serious problem in data analysis because some studies showed that even one of the outliers can affect an entire principle component (PC). This occurs when an extremely high or low value is included in the calculation of mean even it is outliers, thus affecting the overall mean of data and ultimately the covariance matrix, the subsequent eigenvalues and eigenvectors. One study tried to robustify PCA in 1995 by implementing self-organizing rules based on statistical physics approach [21]. By linking the organizing rules to energy function, they proposed an objective function that can resist outliers. However, Xu and Yuille's algorithm has a difficulty to determine the value of hard threshold during the training process. The value will increase to infinity when the value is set to small. Thus, it is almost impossible to find the optimal value of the hard threshold.

Yang and Wang improved Xu and Yuille's algorithm by proposing a fuzzy objective function and gradient descent optimization algorithm that are able to set the value of hard threshold automatically [22]. The new fuzzy objective function has only one parameter which controls the fuzziness variable m that determines the influence of outliers to the weighted average. The higher the fuzziness variable, the sparser and fuzzier the feature space of the clusters will become. Pasi Luukka then published a paper in 2011 where he proposed a nonlinear fuzzy robust PCA which is an improvement from Yang and Wang's objective function by pre-whiten the vector x [23]. The purpose of whitening the vector is that the data will be no longer correlated with each other. The advantage of this approach is that in a tightly clustered data, different attributes will be easily distinguished from one another and the distance between each attribute is more prominent.

In fuzzy PCA, the covariance matrix is determined differently by using fuzzy clustering. FPCA first determines the mean of the attributes by assigning them into a data cluster and noise cluster. The idea is to have a threshold that continuously influences the data by implementing a noise cluster. The center of the cluster will always have the value of zero. Then, the distance of the attributes from the cluster center is calculated in order to obtain a weight for the calculation of the mean. The closer the attributes to the cluster center, the higher the significance of the attributes and given much higher scores. If an attribute's value is extremely high or low, it will be considered as an outlier and will be given lower scores. By utilizing fuzzy membership in PCA, elements with a high degree of membership in cluster center which is a non-outlier will contribute significantly to the weighted average, while outliers with a low degree of membership which are far from the center will contribute almost nothing which will in the end affect the relevancy of each feature. This research utilizes fuzzy membership to rank the features according to its relevancy, while considering outliers, thus

increasing the classification accuracy by removing the issue of bias in classification.

B. SVM steps

Support vector machines is a very popular classification method. It is a supervised learning model with associated learning algorithms that can analyze data and recognize patterns. The performance of SVM is highly dependent on three parameter values that are chosen in the training phase. The first parameter is the Regularization parameter, *c*. Regularization parameter determines the trade-off cost between minimizing the training error and the complexity of the model. The second parameter which is Gamma parameter, *g*, from the kernel function defines the nonlinear mapping from the input space to some high-dimensional feature space. The final parameter is the type of kernel function that is used in the study. Kernel function constructs a nonlinear hyperplane in an input space. In this study, RBF kernel function is chosen and its optimal parameter values are determined by using cross-validation method.

C. Fuzzy c-means

Bezdek introduced Fuzzy c-means clustering method, extended from hard c-mean clustering method [24]. FCM is an unsupervised clustering algorithm that is applied to a wide range of problems related with feature analysis, clustering, and classifier design. FCM is widely applied in agricultural engineering, astronomy, chemistry, geology, image analysis, medical diagnosis, shape analysis, and target recognition [25].

FCM works in the following manners. After the parameters *c* and *m* were entered, FCM calculates the cluster center for each cluster. Each data object has a membership function which determines the degree to which the data object belongs to the certain cluster. Only complete attributes are considered in the process of updating the membership function and centroids. The missing data in the data are determined by utilizing the information about the membership degrees and the values of its cluster centroids. The clustering process stops when the maximum number of iterations (100) is reached, or when the objective function improvement between two consecutive iterations of comparison between the complete dataset and the imputed dataset is less than the minimum amount of improvement specified (0.0001).

3 Proposed model

The proposed model can be divided into two phases which are fuzzy feature selection phase and imputation phase. Here, fuzzy feature selection is implemented through combination of fuzzy PCA (FPCA) and backward sequential selection using SVM, while the imputation phase will be carried out by using FCM. Figure 1 shows the overall flow of the hybrid imputation model, FPCA–SVM–FCM. The above-dashed box indicates the fuzzy feature selection phase by FPCA–SVM, while the below-dashed box indicates the imputation phase.

A. Fuzzy Feature selection phase

In this phase, FPCA acts as filter method that rank the feature based on PC scores, while wrapper approach is used to select the optimum number of features from the dataset. The filter–wrapper hybrid combination of FPCA and backward sequential selection by SVM are used to select relevant and removing irrelevant features from the dataset. Figure 1 shows the proposed hybrid imputation model that consists of fuzzy feature selection with backward sequential selection used to select relevant features to further increase the accuracy of the imputation.

i. Fuzzy principal component analysis (FPCA)

In this phase, FPCA is used to obtain the PC scores for each of the features in order to determine their order of relevancy. The significance of each of the feature is determined by their respective PC scores. The higher the PC scores, the higher the relevancy. In contrast, lower PC scores signify lower relevancy. The most relevant feature will be on top of the rank, while the least relevant features will be at the bottom of the rank.

FPCA is used to identify irrelevant features in the dataset as the irrelevant features could lead to biased result and longer training time. An advantage of FPCA is that data loss is kept at minimum even after the dimension of the data is reduced. FPCA also produces better classification accuracy [23] and faster computational time when compared to classical PCA [20].

ii. Backward sequential selection using SVM

The dataset that has been processed and ranked by FPCA is used as input to SVM. The purpose of this phase is to measure the performance of the selected features. In this study, SVM is employed by backward sequential selection where the classification starts with the complete dataset and begins to delete one by one its features, based on its accuracy performance. The process of selecting features continues until there is no improvement in SVM accuracy. Then, the remaining features are used as input for the imputation method.

SVM classification task involves two processes which are training and testing the data. Each of the data is divided into two parts which are the observed variables (attributes) and its corresponding class labels. The datasets are split into three training–testing partitions, which are 50–50%,

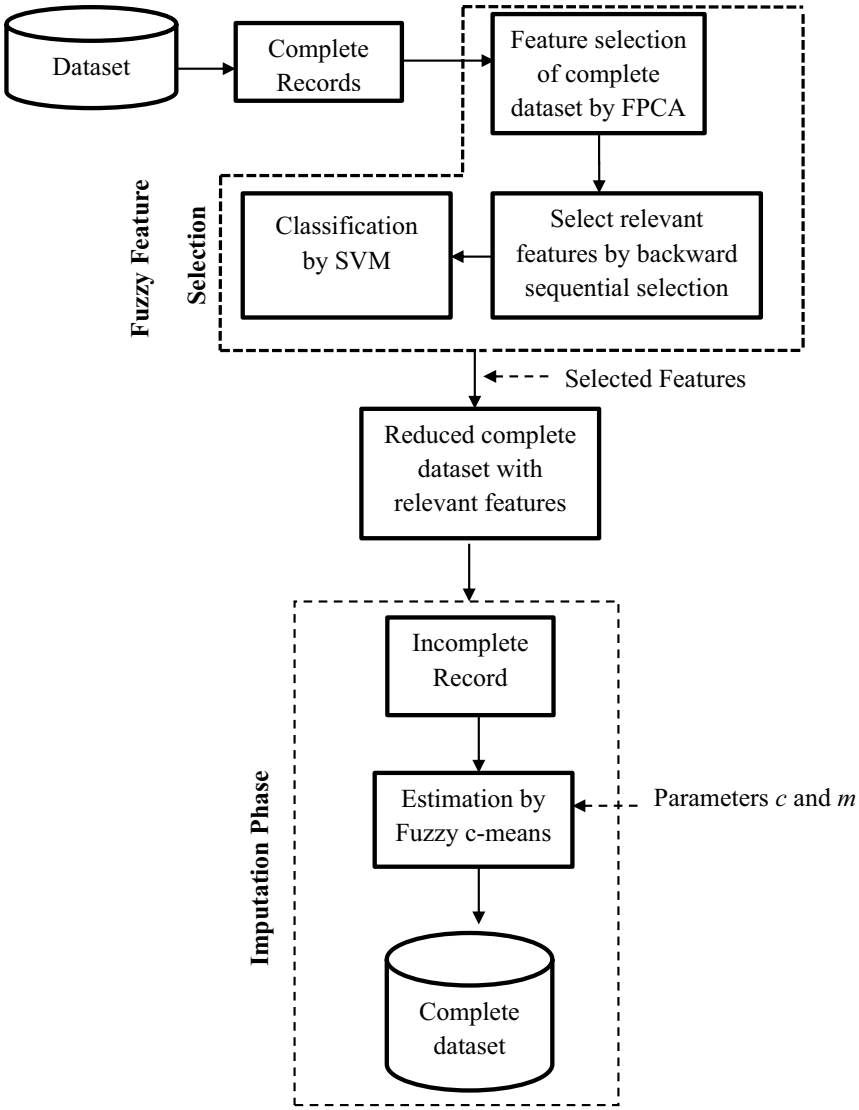

Fig. 1 Flow of hybrid imputation model, FPCA–SVM–FCM

70–30%, and 80–20% to ensure the same class distribution of the subset. The purpose of using different training–testing partitions is to ensure that the imbalance data in the dataset does not affect the overall classification process. The dataset is first trained to obtain the prediction model which is then used to predict the labels during data testing. First, the 50%, 70%, and 80% training sets are trained by using two, five and tenfold cross-validation. After the prediction model is obtained in the training phase by using the best pair of parameters, the dataset undergoes testing phase to predict the class labels. The output from SVM is then processed by FCM to predict the missing data which are artificially introduced in the next phase.

B. Imputation phase

The reduced dataset with the relevant features obtained from FPCA–SVM is used as input to the imputation phase using FCM. The first step in FCM imputation is to artificially delete the completed dataset according to the ratio of 1% up to 50% [26] which is shown in Table 1. The missing data are artificially induced into the dataset to observe the robustness of FPCA–SVM as the missing rate increases [24]. Then, dataset induced with missing data is used in the imputation phase, in which the missing data are estimated by fuzzy c-means.

In order for FCM to work optimally, there are two parameters that are needed to be selected correctly which are *c* and *m*. Here, the parameter *c* and *m* plays a big part. *C* sets cluster number, while *m* sets the weighting factor which controls the fuzziness of the clusters in FCM. The higher the value of *m*, the fuzzier the cluster will become. Data that are far away from the cluster center will be neglected and excluded from the estimation of the missing data. There is no specific value for *c* and *m*. Therefore, in this study, several values that have been proposed in a study

Table 1 Missing data distribution

Dataset	No. of records	Rate of missing data (%)	No. of missing data
Diabetes	768	1	8
		5	38
		10	77
		15	115
		20	153
		25	192
		30	230
		35	268
		40	307
		45	346
		50	384

are used [27]. The value of *c* is 2, 3 or 4, while for parameter *m* it is from 1.5 to 4.

Basically, there are four steps involved in FCM imputation method:

1. Artificially delete some values in the complete dataset obtained from FPCA–SVM to the ratio of 1% up to 50%.
2. Estimate new dataset using FCM.
3. Attain optimized *c* and *m* parameters by using trial–error approach to reduce the error between artificially deleted dataset and complete dataset.
4. Predict the missing data using FCM with the optimized parameters.

To ensure that the best accuracy of imputation has been obtained, two types of measurements are conducted. They are error performance measurement and validation performance. The experiment is then repeated ten times, and its average is calculated. The final output from FCM will be a complete dataset with high accuracy.

4 Results and discussion

In this section, discussion on the result is separated into three parts. The first part will discuss the result obtained from feature ranking by FPCA, while the second part will discuss on the classification result of the selected features by using backward sequential selection. The final and third parts will be on the imputation phase. This section starts with description of the dataset used and performance measurement employed in this study.

a. Experimental data and performance measurement

To verify the effectiveness of the proposed model, a publicly available dataset from the UCI machine learning data repository which is the Pima Indians Diabetes dataset was used. Class 'B' has 458, and class 'M' has 241 instances. Pima Indian Diabetes dataset on the other hand includes a total of 768 instances represented by eight attributes and a predictive class. There are two classes in the dataset which are class "1" and class "0". Class "1" represents diabetic cases, and class "0" indicates non-diabetic cases. Out of 768 instances, 268 instances are in class "1" and the rest of 500 instances are in class "0". In Table 2, the features for all the datasets are shown.

Table 3 shows the partition of data in this study. Here, the dataset is divided into two partitions which are for training and testing with three different percentages of partition. Three percentages of training–testing are employed which are 50–50, 70–30 and 80–20. The purpose of the percentages used is to ensure the same class distribution in the subset. The different training testing percentages also ensure the data imbalance of the dataset did not affect overall classification phase. In the result, the different outcomes from different training–testing partitions are tested and shown.

In this experiment, five performance measurement methods are used to validate the obtained results. The performance measurements were divided into two areas which are for the fuzzy feature selection phase and imputation phase. The performance criterion that is used for evaluating the performance of the fuzzy feature selection

Table 2 Features in the datasets

Diabetes
Body mass index
Diabetes pedigree function
Age
Triceps skinfold thickness
2-h serum insulin
Diastolic blood pressure
Plasma glucose concentration
Number of times pregnant
Predictive class (0–1)

Table 3 Partition of data

Dataset	Training-test partition (%)	No. of records in the subset	
		Training set	Testing set
Diabetes	50–50	384	384
	70–30	538	230
	80–20	614	154

Table 4 Features ranked by FPCA and PCA in diabetes dataset

Method	FPCA	PCA
Feature ranking	1. Body mass index 2. Number of times pregnant 3. Plasma glucose concentration 4. Diastolic blood pressure 5. 2-Hour serum insulin 6. Triceps skinfold thickness 7. Age 8. Diabetes pedigree function	1. Diastolic blood pressure 2. Triceps skinfold thickness 3. Age 4. Diabetes pedigree function 5. Body mass index 6. 2-Hour serum insulin 7. Number of times pregnant 8. Plasma glucose concentration

phase is accuracy. On the other hand, to measure the imputation performance of the proposed hybrid imputation model, FPCA–SVM–FCM, RMSE, and MAE are used to measure the rate of error, while Wilcoxon rank sum and Theil's U test are used to validate the proposed method. To further verify the significance of the proposed hybrid imputation model, FPCA–SVM–FCM, performance comparison with previous studies is conducted. Finally, the best result from each of the table are bolded.

b. Results

i. Features ranking by FPCA and PCA

In this section, how the features are ranked based on the PC scores given for the dataset used in this study is discussed. The proposed method, FPCA is compared with PCA in order to showcase the differences in the features ranking order when outliers are considered. The higher the value of PC scores, the higher the relevancy of the feature compared to the other features. The features are arranged in descending manner which means the first feature is the most relevant feature, while the last feature is the least significant feature. The dashed line signifies the cutoff point where the features ranked below the dashed line are deleted and considered irrelevant in the classification phase later.

1. Pima Indians diabetes dataset

Next, Pima Indians Diabetes dataset is processed by FPCA and PCA. The results from the features selected are shown in Table 4. FPCA and PCA both identified different features as the most relevant features. FPCA identified "BMI" as the most relevant feature, while PCA identified "Diastolic blood pressure". Both methods also select different features as the least relevant features which are "Diabetes pedigree function" and "Plasma glucose concentration" by FPCA and PCA, respectively.

ii. Backward Sequential Selection Result Using SVM Classification:

In this phase, the classification performance of FPCA–SVM with PCA–SVM and SVM is presented. The classification performance of the original dataset by SVM is also noted to obtain the benchmark performance with the irrelevant features which are still present. Below are the results of classification accuracy. Classification accuracy is used as the performance criterion to remove irrelevant features and to obtain optimum number of relevant features. To further validate the effectiveness of the proposed model, comparisons with the previous studies have also been carried out.

1. Pima Indians diabetes dataset

Classification results of the Pima Indians diabetes dataset for PCA–SVM and SVM are collected at Table 5. The method is listed in the first column; the second column gives the result of classification accuracy by each partition of the data. In the table, it is shown that by reducing the dimension with FPCA, the classification accuracy increases in comparison with the classification by SVM.

As can be seen from Table 5, FPCA–SVM classifies diabetes with the accuracy of 72.078% using four features, whereas PCA–SVM produces the highest classification accuracy with 69.286% using also four but different sets of features in comparison with FPCA–SVM. There are noticeable increases in the classification accuracy, especially in the 70–30 partition with about 5% of increase. In the 80–20 partition, there is a slight improvement of 3% increase when compared to the SVM and PCA. Although the increase of 3% in classification accuracy is not so big, any increase in classification accuracy is a good indication. In

Table 5 Comparative classification results between FPCA, PCA, and SVM for diabetes dataset

Method	Classification accuracy (%)		
	50–50	70–30	80–20
$Diabetes_{SVM}$	68.052	65.652	69.286
$Diabetes_{PCA}$	68.052	65.652	69.286
$Diabetes_{FPCA}$	**70.052**	**70.652**	**72.078**

the Pima Indian Diabetes dataset, fuzzy c-means produced the highest classification accuracy when the parameters for *c* and *m* are 3 and 1.5, respectively. Indeed, selecting the correct required features can increase the classification accuracy.

As the table shows, SVM and PCA have the same classification accuracy. This is probably because PCA ranks relevant features lowly, thus not contributing to the increase in the classification accuracy. The proposed method, FPCA–SVM with the fuzzy element can further increase the accuracy by finding the most significant features as shown in the result. FPCA ranks "Diabetes pedigree function", "Age" and "Triceps skin fold thickness" among the least relevant features, while PCA ranks the three features highly. PCA also ranks important features such as "Plasma glucose concentration" and "Number of times pregnant" as irrelevant features, thus reducing the classification accuracy. This shows that FPCA can rank the features better than classical PCA.

iii. Imputation phase:

After the set of relevant features were determined, the reduced dataset is then artificially induced with missing data with certain percentage to determine the robustness of our method with different rates of missing data. The missing data will be estimated by FCM. The imputation performance evaluation is based on four performance measurement methods which are RMSE, MAE, Wilcoxon rank sum, and Theil's U test. The performances were evaluated according to their respective missing data rates which are 1% up to 50%. Each of the evaluation is repeated ten times and calculates its average in order to obtain a more comprehensive result. The proposed model also was compared to SVM–FCM to observe whether the inclusion of feature selection has an effect.

1. RMSE

RMSE calculates the error between the real values and the estimated values (imputed values) to measure the accuracy of the imputation. RMSE provides valuable insight into the short-term performance of the method by comparing the differences between true value and estimated values. The performance of our proposed model also improved compared to SVM–FCM method as shown in Table 6. This is shown in the table where our method has lower RMSE compared to SVM-FCM at every missing data rate. Even with the increasing missing rates up to 50%, the proposed method performed better than SVM-FCM. This is because FPCA reduces the number of dataset dimensions compared to the full dataset [28], thus allowing FCM to perform better.

Table 6 Comparison of RMSE between the proposed model and SVM–FCM for diabetes datasets

RMSE value											
1%		5%		10%		15%		20%		25%	
FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM
0.003	0.007	**0.016**	0.017	**0.027**	0.027	**0.03**	0.035	**0.049**	0.050	**0.063**	0.075
30%		35%		40%		45%		50%			
0.077	0.088	**0.080**	0.094	**0.085**	0.115	**0.091**	0.139	**0.098**	0.148		

2. MAE

From Table 7, we can see that the average of errors between actual value and estimated value are minimal. The MAE results we obtain also accord with RMSE. This is important because both are error-based performance measure metrics. This shows that our proposed model produces minimal error compared to SVM–FCM. Although the MAE value increases significantly with the increasing missing rates, the result can be considered good when compared with SVM–FCM and other studies which are presented in the next section. Another reason for improvement is that by utilizing FPCA–SVM before the imputation process, the outliers that could affect the calculation of the missing data are removed.

3. Wilcoxon rank sum

A Wilcoxon rank sum test is a nonparametric test that can be used to determine whether two independent samples were selected from populations having the same distribution. Wilcoxon rank sum is interpreted by looking at the *P* value. Higher P value signifies a more accurate estimate rather than lower *P* values. In Table 8, we can see that mostly the result returns a high P value which means a very accurate prediction by our proposed model. Even at the 50% missing rates, the Wilcoxon rank sum value does not drop lower than 0.5 which is the point where the result can be considered bad. This demonstrates the importance of feature selection in increasing the predictive performance of FCM.

4. Theil's *U* test

The final performance measurement for the imputation result by using Theil's *U* test is presented in Table 9. *U* test is a relative accuracy measure that compares the forecasted results with the results of forecasting with minimal historical data. It also squares the deviations to give more weight to large errors and to exaggerate errors, which can help eliminate methods with large errors. If the *U* value is lower than 1, it indicates greater predicting accuracy, while *U* value more than 1 indicates otherwise. In this study, FCM produces near to 0 *U* value which further solidifies the performance of FCM. Even when the missing data present are half of the whole dataset, the proposed method produced good Theil's *U* test value at 0.098. This shows that FPCA–FCM is very robust even when the missing data present are high.

Table 7 Comparison of MAE between the proposed model and SVM–FCM for diabetes datasets

MAE Value													
1%		5%		10%		15%		20%		25%			
FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM	FPCA–FCM	SVM–FCM		
0.001	0.003	**0.001**	0.002	**0.001**	0.001	**0.003**	0.003	**0.005**	0.005	**0.008**	0.009		
30%		35%		40%		45%		50%					
0.011	0.012	**0.013**	0.026	**0.0143**	0.022	**0.0214**	0.0289	**0.0282**	0.0354				

Table 8 Wilcoxon rank sum value imputation validation result for diabetes datasets

Wilcoxon rank sum value					
1%	5%	10%	15%	20%	25%
0.999	0.959	0.797	0.674	0.606	0.599
30%	35%	40%	45%	50%	
0.584	0.569	0.513	0.498	0.496	

Table 9 Theil's *U* test imputation validation result for diabetes datasets

Theil's *U* value					
1%	5%	10%	15%	20%	25%
0.002	0.004	0.005	0.010	0.012	0.031
30%	35%	40%	45%	50%	
0.049	0.058	0.062	0.073	0.098	

Table 10 Comparison of Wilcoxon rank sum value with previous methods in diabetes dataset

Method	Wilcoxon rank sum
FPCA–SVM–FCM	**0.797**
PSO_COV	0.73
K-Means + MLP	0.73

4.1 Comparative analysis

The performance of FPCA–SVM–FCM is further validated by comparing the performance with several published methods that used the same dataset.

As shown in Table 10, FPCA–SVM–FCM performs better than the previously published method of PSO_COV and K-Means plus MLP [29]. The higher the Wilcoxon rank sum value, the better the performance of the methods. This indicates that using a feature selection before imputation phase can increase the accuracy of imputation considerably. Irrelevant features can introduce bias thus reducing the accuracy of imputation. By using FPCA–SVM, the irrelevant features are deleted and only irrelevant features are kept.

The proposed model is also compared with another study that was applied to the diabetes dataset. The method used by previous author was Gray Fuzzy Neural Network (GFNN) [30]. It works by using optimal parameters obtained from Gray Wolf Optimizer (GWO) to optimize the membership function and then impute the data for both categorical and numerical data by using Adaptive Neuro-Fuzzy Inference System (ANFIS). In the paper, the GFNN was compared to several previous methods such as KNN, WLI, and GWLMN at 20% rate of missing data. The results are shown in Table 11.

Table 11 Comparison of RMSE with previous methods in diabetes dataset

Method	RMSE (20% Missing rate)
FPCA–SVM–FCM	**0.049**
KNN	26.20
WLI	8.826
GWLMN	8.611
GFNN	4.930

As shown in Table 11, it clearly indicated that FPCA–SVM–FCM produce much lower RMSE value compared to the rest of the methods above. GFNN produces RMSE value of 4.930, while our proposed method produces much lower RMSE at 0.049. This might due to the fact that the GFNN never considers the presence of outliers in the dataset. Although GFNN operates at optimum parameters, the presence of outliers can skew the predicted value form ANFIS. Outliers in the dataset can influence the outcome of imputation as most imputation methods predict the missing values are based on the remaining values in the dataset and one study showed that even one outlier can affect the result obtained. Thus, it proved that a feature selection method that can consider the presence of outliers was able to increase the imputation accuracy of the imputation method.

Finally, we compare our proposed method with a study that imputed 84 real-world datasets taken from the UCI Repository which include the dataset that is used in our study which is the Pima Indians diabetes dataset. The study benchmarked their proposed methods which are opt.knn, opt.svm, and opt tree against existing imputation methods including mean impute, K-nearest neighbors (KNN), iterative knn (iKNN), Bayesian PCA (BPCA), predictive-mean matching (PMM) [31]. All three proposed methods by the author are the product from optimizing the missing values in all data points and dimensions simultaneously by using K-nearest neighbors, SVM, and decision tree-based imputation. The results are tabulated in Table 12.

The methods are compared by using MAE with 30% of missing data. In the table, we demonstrate that the data imputation predicted by our proposed method gives much better performance compared to the rest of the benchmark methods. Again, all the methods in Table 12 did not consider the influence of outliers thus reducing its accuracy. This further proved that by using our proposed method while considering outliers, the predictive performance of the imputation methods can be increased significantly.

Table 12 MAE comparison of benchmark methods, Opt. impute methods, and FPCA–SVM–FCM in 30% of missing data rate

Methods	MAE
FPCA–SVM–FCM	**0.011**
Mean	0.1217
PMM	0.1453
BPCA	0.1109
KNN	0.1164
iKNN	0.1098
Opt.knn	0.1098
Opt.svm	0.1049
Opt.tree	0.1069

5 Conclusion

Missing data in medical dataset can introduce the issue of accuracy and bias when creating a diagnosis or conclusion from a case files. Thus, there is a need to develop a good imputation method that can predict the missing data with high accuracy. However, the presence of outliers in datasets can reduce the effectiveness of the imputation method as extremely high value will affect the calculation of the missing data. Outliers can also render some of the features that become irrelevant. One of the best known methods to remove irrelevant features is by using a feature selection method.

In this paper, a new hybrid imputation method, FPCA–SVM–FCM has been proposed. Here, the feature selection method used is Fuzzy Principal Component Analysis (FPCA) where it identifies relevant features in dataset with the consideration of outliers. Support Vector Machines is then used to classify the selected features and delete irrelevant features. After the significant features in the dataset are identified, the missing data are then imputed by Fuzzy c-means (FCM).

Experimental results that are applied on one medical dataset which is the Pima Indians Diabetes datasets show that FPCA–SVM has produced a substantial increment in classification performance for the dataset compared to classical PCA and SVM in terms of accuracy. Fuzzy membership in PCA has helped to increase the capability of PCA in recognizing significant feature correctly. This is due to the capability of FPCA to differentiate attributes as outliers by dividing the feature space using distinct approach from classical PCA. Therefore, FPCA can produce better learning and generalization ability in SVM classifier. By removing the irrelevant features, the imputation performance of FCM performs well in terms of RMSE, MAE, Wilcoxon rank sum, and Theil's U test when compared to SVM–FCM. The increase in FCM performance is due to no presence of outliers that affect the calculation of the missing data.

It is believed that the promising results demonstrated by FPCA–SVM–FCM can be used to assist medical practitioners in the healthcare practice for better and precise diagnosis. Future work will focus on optimizing the parameters of the methods as three methods used in this research have multiple parameters that are needed to be chosen systematically.

Acknowledgement This study is supported by the Fundamental Research Grant Scheme (FRGS vot: 4F738) that was sponsored by Ministry of Higher Education (MOHE). Authors would like to thank Research Management Centre (RMC) Universiti Teknologi Malaysia, and Soft Computing Research Group (SCRG) for the support in research activities.

Compliance with ethical standards

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

References

1. Lang KM, Little TD (2018) Principled missing data treatments. Prev Sci 19(3):284–294
2. Che Z, Purushotham S, Cho K, Sontag D, Liu Y (2018) Recurrent neural networks for multivariate time series with missing values. Sci Rep 8(1):6085
3. Yan X, Xiong W, Hu L, Wang F, Zhao K (2015) Missing value imputation based on gaussian mixture model for the internet of things. Mathematical Problems in Engineering
4. Basak D, Pal S, Patranabis DC (2007) Support vector regression. Neural Inf Process-Lett Rev 11(10):203–224
5. Panigrahi L, Das K, Mishra D (2014) Missing value imputation using hybrid higher order neural classifier. Indian J Sci Technol 7(12):2007
6. Pan R, Yang T, Cao J, Lu K, Zhang Z (2015) Missing data imputation by k nearest neighbours based on grey relational structure and mutual information. Appl Intell 43(3):614–632
7. Jörnsten R, Wang HY, Welsh WJ, Ouyang M (2005) DNA microarray data imputation and significance analysis of differential expression. Bioinformatics 21(22):4155–4161
8. Aydilek IB, Arslan A (2013) A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Inf Sci 233:25–35
9. Dai LY, Feng CM, Liu JX, Zheng CH, Yu J, Hou MX (2017) Robust nonnegative matrix factorization via joint graph Laplacian and discriminative information for identifying differentially expressed genes. Complexity
10. Meesad P, Hengpraprohm K (2008) Combination of knn-based feature selection and knnbased missing-value imputation of microarray data. In: Innovative computing information and control. ICICIC'08. 3rd International conference on (pp 341-341). IEEE (2008)
11. Doquire G, Verleysen M (2012) Feature selection with missing data using mutual information estimators. Neurocomputing 90:3–11
12. Shi X, Guo Z, Nie F, Yang L, You J, Tao D (2016) Two-dimensional whitening reconstruction for enhancing robustness of principal component analysis. IEEE Trans Pattern Anal Mach Intell 38(10):2130–2136
13. Howard WJ, Rhemtulla M, Little TD (2015) Using principal components as auxiliary variables in missing data estimation. Multivar Behav Res 50(3):285–299

14. Huang X, Maier A, Hornegger J, Suykens JA (2017) Indefinite kernels in least squares support vector machines and principal component analysis. Appl Comput Harmon Anal 43(1):162–172
15. Xu J, Yin Y, Man H, He H (2012) Feature selection based on sparse imputation. In: Neural networks (IJCNN), the 2012 international joint conference on (pp 1–7). IEEE
16. Koutanaei FN, Sajedi H, Khanbabaei M (2015) A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. J Retail Consum Serv 27:11–23
17. Purnami SW, Rahayu SP, Embong (2008). A feature selection and classification of breast cancer diagnosis based on support vector machines. In: Information technology, 2008. ITSim 2008. International symposium on (vol 1, pp 1–6). IEEE
18. Shen F, Shen C, Liu W, Tao Shen H (2015) Supervised discrete hashing. In: Proceedings of the IEEE conference on computer vision and pattern recognition 37–45
19. Akay MF (2009) Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst Appl 36(2):3240–3247
20. Gharibnezhad F, Mujica Delgado LE, Rodellar Benedé J, Fritzen CP (2013) Damage detection using robust fuzzy principal component analysis. In: Proceedings 6th European workshop on structural health monitoring (pp 1–6)
21. Xu L, Yuille AL (1995) Robust principal component analysis by self-organizing rules based on statistical physics approach. IEEE Trans Neural Netw 6(1):131–143
22. Yang TN, Wang SD (1999) Robust algorithms for principal component analysis. Pattern Recognit Lett 20(9):927–933
23. Luukka P (2011) A new nonlinear fuzzy robust PCA algorithm and similarity classifier in classification of medical data sets. Int J Fuzzy Syst 13(3):153–162
24. Bezdek JC (1974) Numerical taxonomy with fuzzy sets. J Math Biol 1(1):57–71
25. Yong Y, Chongxun Z, Pan L (2004) A novel fuzzy c-means clustering algorithm for image thresholding. Meas Sci Rev 4(1):11–19
26. Purwar A, Singh SK (2015) Hybrid prediction model with missing value imputation for medical data. Expert Syst Appl 42(13):5621–5631
27. Wu KL (2012) Analysis of parameter selections for fuzzy c-means. Pattern Recognit 45(1):407–415
28. Michalak K, Kwasnicka H (2010) Correlation based feature selection method. Int J Bio-Inspired Comput 2(5):319–332
29. Krishna M, Ravi V (2013). Particle swarm optimization and covariance matrix based data imputation. In: Computational intelligence and computing research (ICCIC), 2013 IEEE international conference on (pp 1–6). IEEE
30. Kuppusamy V, Paramasivam I (2017) Grey fuzzy neural network-based hybrid model for missing data imputation in mixed database. Int J Intell Eng Syst 10:146–155
31. Bertsimas D, Pawlowski C, Zhuo YD (2017) From predictive methods to missing data imputation: an optimization approach. J Mach Learn Res 18:1–196