Feature Based Multivariate Data Imputation

Alessio Petrozziello and Ivan Jordanov

School of Computing, University of Portsmouth, UK
alessio.petrozziello@port.ac.uk, ivan.jordanov@port.ac.uk

Abstract. We investigate a new multivariate data imputation approach for dealing with variety of types of missingness. The proposed approach relies on the aggregation of the most suitable methods from a multitude of imputation techniques, adjusted to each feature of the dataset. We report results from comparison with two single imputation techniques (*Random Guessing* and *Median Imputation*) and four state-of-the-art multivariate methods (*K-Nearest Neighbour Imputation*, *Bagged Tree Imputation*, *Missing Imputation Chained Equations*, and *Bayesian Principal Component Analysis Imputation*) on several datasets from the public domain, demonstrating favorable performance for our model. The proposed method, namely *Feature Guided Data Imputation* is compared with the other tested methods in three different experimental settings: *Missing Completely at Random*, *Missing at Random* and *Missing Not at Random* with 25% missing data in the test set over five-fold cross validation. Furthermore, the proposed model has straightforward implementation and can easily incorporate other imputation techniques.

Keywords: *Missing data; Multivariate data imputation; Multitude of imputation models; Data mining*

1. INTRODUCTION

Dealing with missing data is an important step in dataset pre-processing since most statistical analysis techniques, data reduction tools, and machine learning methods require complete datasets. There are many techniques that can be used to deal with the missingness, but the common approach during imputation is to make the most of the available data through minimizing the loss of statistical power and the bias inevitably brought by the missing data inferred values. The mechanisms of missingness are usually categorized into three groups [1]: MCAR (Missing Completely at Random); MAR (Missing At Random); and MNAR (Missing Not At Random). In the first case, the missingness is generally due to external factors, not correlated to the other variables in the dataset, while in the last two, the cause is related to the other variables; therefore, the risk of bringing bias due to the imputation should be carefully considered.

The approaches of dealing with missingness can be also divided into three categories [1]: deletion; univariate imputation; and multivariate imputation. In the first category fall the list-wise deletion (the patterns with missing values are simply removed), attribute deletion (the features with missing values are excluded) and pairwise deletion (where, in presence of missing values, the pattern is not dropped, and its other values are still used during the analysis). The methods from the second category do not consider the correlation between the missing value and the other variables in the dataset, and impute the data using only information of the same attribute. Good examples of this group are: the *Random Guessing*,

where the values are substituted randomly, sampling from the other values of the same attribute; and the *Mean* (*Median*) *Imputation*, where the values are replaced with the mean (median) of the considered attribute. The last category includes methods that consider the correlation of the different attributes. Four different algorithms of this family are usually considered [1]: *Multiple Imputation Chained Equations* (MICE); *Bagged Tree Imputation* (BTI); *K-Nearest Neighbour Imputation* (KNNI) and *Bayesian Principal Component Analysis Imputation* (bPCA).

These methods have been widely investigated and compared in the past years, showing discordant results [2] [3]. Most approaches of dealing with missingness would select a single method that outperforms the others based on a given performance metrics. However, while a given approach might have a good performance across the whole dataset, it does not mean that its performance will be superior at the level of each individual feature. In the proposed approach, instead of selecting a single method which outperforms the others on the whole dataset, a column-wise selection is used to choose the best imputation method for each individual attribute.

The proposed method, namely *Feature Guided Data Imputation* (FGDI) is extensively tested and validated on thirteen publicly available datasets. Its performance is assessed and compared with other techniques using *Wilcoxon Signed-rank* test for statistical significance [4].

The remainder of the paper is organized as follows. Section 2 describes the considered imputation methods, while Section 3 proposes the FGDI method. Section 4 discusses the empirical study carried out. The results of this investigation are discussed in Section 5 and in Section 6 conclusion given.

2. IMPUTATION TECHNIQUES

Baselines - The most common techniques used as baselines for comparison and analysis of data imputation are *Random Guessing, Mean Imputation* and *Median Imputation* [5]. The *Random Guessing* is a very simple benchmark to estimate the performance of a prediction method. It takes as input the missing data with random value drawn from the known values of the same feature. The *Mean* (*Median*) *Imputation* replaces every missing value with the mean (median) of the attribute. However, these techniques fall into the single imputation category (the correlation between the variables is ignored), which is the reason for being rejected by the scientific community [6], hence, they are only used here to perform initial fast sanity check of the proposed approach.

Bagged Tree Imputation - The BTI with gradient boosting [7] is a machine learning technique for solving regression problems, which produces a robust prediction model using a vote (ensemble) among weak ones. The method follows few basic steps for each feature with missing data: (1) train several tree models using the other features; (2) for each tree, impute the data using a regression function; (3) use a vote among the trees to select the data that will be imputed in the original dataset. Bagging predictors are used for generating multiple versions of a predictor to get an aggregated one. The aggregation uses the average over the predictor versions when predicting a numerical outcome, and employs a plurality

vote when predicting a class. Bagging proved to be more efficient in the presence of label noise when compared to boosting and randomization [8]; it is also robust to outliers and can impute the data very accurately using surrogate splits [9]. Another important feature of the tree model is its flexibility: different models can be trained with the random forests and the prediction deferred to a system vote among them. In this work, we employ gradient boosting technique for the regression values, which uses an ensemble of weak decision trees.

K-Nearest Neighbors Imputation – In the KNNI the missing values are usually imputed applying the mean of the *K* most similar patterns found by minimizing the *Euclidean Distance* between a pattern with missing values and the complete subset [10]. The KNNI approach comprises three steps: (1) take only the rows of the dataset without missing data and use this subset as a prototype dataset to select the nearest neighbours; (2) choose a distance metric and compute the nearest neighbours between each pattern with missing data and the complete subset; (3) impute the data, using the mean or the mode of the chosen neighbours. An important parameter to select is the number of neighbours *K*. There are discordant opinions in the literature, some suggesting a low value of 1 or 2 for small datasets [11]. [12] advise a value of 10 for large datasets, and in [10] is argued that the method is insensitive to the choice of the number of neighbours. In all simulations carried out in this work, we used a value of $K = 10$. The K-Nearest Neighbours has some advantages: the method can predict both, categorical variables (the most frequent value among the KNN) and continuous variables (the average among the KNN); and when using this imputation, there is no need to build a model (as in the Bagged Tree Imputation).

Missing Imputation Chained Equations - MICE [13] is a method from the multiple imputation family. In the MICE process, a series of regression models are run modeling each variable with missing data as dependent variable relying on all the other variables in the dataset. This guarantees that each variable is modeled independently to its distribution [13]. The MICE method is divided into four stages: (1) a simple imputation (Mean) is performed for every missing value in the dataset to be used as placeholders; (2) the placeholders for one variable are set back to miss; (3) the missing variable is used as the dependent variable in a regression model and regressed using the other variables. The procedure is followed for every variable with missing entries and repeated many times until the convergence is reached. Practical guide on how to select the number of imputations is given in [14], however, sometimes due to the size of the dataset, it is not feasible to run the procedure many times. Therefore, 10 iterations are usually considered enough for the convergence of the algorithm [15], which number is also adopted in this investigation.

Bayesian Principal Component Analysis Imputation - the bPCA imputation [16] is an evolution of the Single Value Decomposition Imputation [10] (since the SVD is a PCA applied to normalized datasets with a 0 row-mean) with the additional Bayesian estimation, using a known prior distribution. An advantage of this approach is that no hyper-parameters tuning is needed, and the number of components is self-determined by the algorithm at the expense of a higher computational time. The bPCA can be summarized as: (1) apply Principal Component Regression on the initial dataset; (2) perform a Bayesian Estimation; (3) use an EM algorithm until convergence to a specified tolerance.

3. THE PROPOSED METHOD

All methods described in the previous section have been widely applied for solving missing data problems [2]. However, while a given approach may produce low estimation error for the whole dataset at hand, this does not mean that the method outputs the best result (smaller error) for every individual feature (usually, for some of the features other methods may give better estimates).

The investigated here *Feature Guided Data Imputation* (FGDI) is an imputation approach which aggregates models in a feature-wise fashion (choosing the best model for each feature (column) of the dataset, while allowing it at the same time to be inferior for the rest of the features). In other words, when training the model, the best imputation method for each feature of the dataset is selected among the considered techniques. At the imputation phase, each selected method is sequentially used to impute the features for which its performance was the best during the training stage.

During the learning phase, the algorithm is trained on artificially introduced missing data (e.g., 25% of MCAR, MAR or MNAR) for each feature. A combination of the best performed methods (based on a given error metrics, e.g., RMSE, MAE) is used to impute the missing values in the original dataset. To cope with the random nature of the algorithm and to ensure more robust choice, this process is iterated a given number of times, and the technique that produced the lowest median overall error for each feature is then chosen. For example, let's assume a set of m imputation methods ($M_1, ..., M_m \in S$) and dataset (X) composed of v variables (features) and n samples, where k of them ($0 < k < n$) contain at least one missing value. Once the n-k complete samples are separated (X' subset), a percentage of missingness is added to each variable of X' (e.g., 25%). The missing data in X' are separately imputed using all methods of S, and the estimation error (e.g., RMSE) is calculated for each feature (variable). This process is repeated I times (e.g., $I = 5$), and for every variable in X', the imputation algorithm scoring the lowest median error is selected and included in a set E, ($E \subseteq S$).The selected techniques are then used to estimate the missing values of the whole set X. In particular, $\forall M_i \in E, i = 1,.., j$, (where $j \leq m$), the dataset X is entirely imputed, and only the imputed values for the features where M_i scored the lowest error are saved, discarding the others. Since X is imputed independently using each technique, the order of imputation is irrelevant, enabling the process to be parallelized.

4. EMPIRICAL STUDY

In previous works [3] [17], extensive review and experimentation was done in an effort to identify correlation between imputation methods performance and the type of datasets with missingness, which concluded with discordant results (confirming the 'No free lunch theorem'). These findings led to the current investigation, based on the aggregation of different models.

The proposed method (FGDI) is compared with known univariate baselines and multivariate state-of-the-art imputation methods (i.e., KNN, BTI, MICE and bPCA) to assess its performance on the missing data imputation task. The experiments are executed for all

the three missing data mechanisms: MCAR, MAR and MNAR. Lastly, a run time analysis is carried to observe the computational cost needed during the training and imputation phases. The results are reported in Section 5.

Thirteen publicly available datasets from KEEL repositories [18] are used in this work, namely *Contraceptive, Yeast, Red wine, Car, Titanic, Abalone, White Wine, Page Block, Ring, Two Norm, Pen Based, Nursery*, and *Magic04*. The selection of these datasets was driven by the intent to cover different application domains and data characteristics. They differ in the number of instances (from several hundreds to several thousands), the number of features (from 3 to 20), and in the range and type of the features (real, integer and categorical). The used datasets do not have missing values by default, guaranteeing total control over the experiments and the assessment and evaluation of the results.

From the variety of metrics employed for comparing and evaluating data imputation and prediction models found in the literature, *Mean Squared Error* (MSE) and *Mean Absolute Error* (MAE) are the most widely used [16] [19]. *MSE* measures the difference between predicted and actual values while MAE their absolute difference. The *Mean Absolute Error* (MAE) is argued to be more accurate and informative than the RMSE [20], successively refuted by [21], where it is stated that the two measures picture different aspects of the error and therefore they should both be used to assess the results. As suggested in [20] and [21], RMSE and MAE are implemented to compare the estimated missing values and the original ones, reflecting the average performance of the imputation method. Furthermore, the RMSE is employed as error function for the training phase of the FGDI. The *Standard Accuracy* (SA) and *Variance Relative Error* (RE*) are assumed to be good baseline estimation measures [22]. SA and RE* are used to compare the proposed model with the univariate baseline imputation techniques (discussed earlier). In particular, SA which compares the prediction against the mean of a random sampling of the training response values $SA=1 - RMSE(predicted, actual) / RMSE(randGuess, actual)$ and the $RE^*=\sigma^2(predicted - actual) / \sigma^2(actual)$ which gives score of 1 for a model predicting values with 0 variance. It is considered an appropriate baseline error measure since any model producing RE* greater than 1 would be assumed weak, independently of the dataset [22].

To validate the proposed method, a k-fold cross validation is applied, splitting the dataset into independent training and test sets. The test set is generated using a uniform sampling without repetitions, and the rest of the data is left as a training set. Since the *Shapiro Test* showed that many of our patterns came from non-normally distributed populations, the statistical *Wilcoxon Signed Rank Test* was used to prove which method is giving better performance [4]. Furthermore, the used test does not make any assumptions about the underlying distribution of the data. In order to check the statistical significance of the difference in model performance, we test the following *NULL* hypothesis: "Given a pair of models (M_i, M_j) with $i, j \in \{1, .., n\}, i \neq j$, the RMSEs (MAEs) obtained by model M_i are significantly smaller than the errors produced by model M_j", using confidence level $\alpha=0.05$.

When simulating *Missing Completely at Random* (MCAR) mechanism, for each feature value in the dataset, a number is drawn from a uniform distribution in the (0, 1) interval. If this number is smaller than assumed missing data threshold (e.g., 0.25), the feature value is set as missing in the original dataset. For the *Missing at Random* (MAR) mechanism, a

variance-covariance matrix is built for the considered dataset. For each variable, the probability of missingness is governed by the most correlated feature in the matrix (i.e., the bigger the value of the correlated feature, the higher the probability of introducing missingness). To generate the *Missing Not at Random* (MNAR) mechanism, we draw values (used as thresholds) from a uniform distribution in (0, 1) interval, and sort them in decreasing order. We do the same for the variable values and pair them with the sorted random numbers. For each threshold, we draw a new random number in the (0, 1) interval and if it is smaller than the threshold, we erase the feature value (this way the pairs with higher random numbers are more likely to be set as missing).

5. RESULTS AND DISCUSSION

Three different experiments are carried out: MCAR, MAR, and MNAR mechanisms with 25% of missing data and 5-fold cross validation (80% training and 20% testing). To calibrate the model during the training phase, 25% of missing data is added to each attribute of the training set, subsequently imputed using the five imputation techniques and the accuracy is evaluated using both MAE and RMSE. This process is run 5 times and for each attribute, the imputation model achieving the lowest median error (preferred to the mean due to robustness to outliers) is selected. Lastly, the selected techniques are used to impute the data on the independent test set and the results are compared to all the other methods.

The first set of experiments is performed imputing the missing data under the MCAR mechanism. As the MCAR occurs when the missingness is unrelated to anything in the study, the missingness is simulated using a Bernoulli random variable removing values with 25% chance of success. The *SA* values given in Table 1 show superior results for the imputation carried out with our model. It outperformed the baseline methods *Random Guessing* (SA_{Random} *is always 0*) and the *Median Imputation* ($SA_{FGDI} > SA_{Median}$). The *Mean Imputation* was omitted in favor of the *Median Imputation*, since the latter is considered less biased to outliers. Furthermore, Table 2 presents the *RE** results over five different imputation methods and again, as it can be seen from the values, our FGDI method outperformed the *Median Imputation*, with $RE_{FGDI} < 1$ in almost all case studies. It can be also seen from the table that the $RE_{MICE} > 1$, which means high variance in the imputed values, problem already discussed in [23]. The RE_{KNNI}, instead, shows high variance (from 0.19 to 1.24) depending on the considered dataset and feature. In the *Yeast* dataset, two variables (*Erl* and *Pox*) are removed during the RE* calculation since the variance in the denominator is 0. To finally assure that the proposed method is outperforming the baselines, a *Wilcoxon* test for statistical significance is run, testing the *NULL hypothesis* "The RMSEs provided by FGDI are significantly smaller than the errors produced by the models *Random Guessing* and *Median Imputation*". The results proved FGDI being better than both with $p\text{-}value < 0.05$ over all 13 datasets. The *Standard Accuracy* analysis (Table 1) shows that the FGDI method not only outperforms the baselines, but it is also comparable, and even better than the state-of-the-art algorithms. As it can be seen from the table, the SA_{FGDI} is higher than the *SA* of the other methods in 41 out of the 52 cases, comparable in 9 out of the 52 cases, and worse in only 2 cases. To validate the significance of the difference, the *Wilcoxon* test is run justifying the *NULL hypothesis* "The RMSEs provided by FGDI are significantly smaller than the errors achieved by the state-of-the-art methods".

Table 1: Standard Accuracy (SA) values achieved by FGDI, the baseline (Median Imputation) and state-of-the-art (KNNI, BTI, MICE, and bPCA) techniques over the 13 datasets for 5-fold cross validation with 25% MCAR. Higher values represent better estimation over the random guess

Dataset	FGDI			KNNI			BTI			MICE			bPCA			Median		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
Contraceptive	**0.39**	**0.27**	**0.31**	0.24	0.11	0.17	0.36	0.26	0.30	0.18	-0.02	0.03	0.38	0.23	**0.31**	0.26	0.23	0.27
Yeast	**0.33**	**0.37**	**0.27**	0.24	0.29	0.09	0.32	0.36	0.24	0.06	0.04	-0.02	**0.33**	0.36	0.22	0.28	0.36	0.22
Red Wine	**0.37**	**0.28**	**0.28**	0.33	0.13	0.14	0.33	0.18	0.25	0.23	0.01	-0.08	0.32	0.15	0.25	0.30	**0.28**	0.26
Car	**0.29**	**0.32**	**0.29**	0.12	0.21	0.16	**0.29**	0.31	0.15	-0.01	0.01	-0.07	**0.29**	0.31	0.14	0.25	0.29	**0.29**
Titanic	**0.35**	**0.27**	**0.28**	0.26	0.18	0.00	0.34	**0.27**	0.26	0.05	-0.06	-0.05	0.34	**0.27**	0.23	0.28	0.25	0.26
Abalone	0.68	**0.28**	**0.27**	0.62	-0.32	-0.05	0.57	0.27	0.18	0.66	0.08	-0.10	**0.72**	0.08	-0.10	0.28	**0.27**	**0.27**
White Wine	**0.36**	**0.29**	**0.30**	0.34	0.11	0.12	0.34	0.18	0.18	0.16	-0.01	0.00	0.34	0.18	0.19	0.28	**0.29**	0.29
Page Block	**0.49**	**0.26**	0.22	0.41	0.16	0.17	0.43	**0.26**	0.20	0.39	0.12	0.03	0.46	0.22	0.16	0.25	**0.26**	**0.23**
Ring	**0.31**	**0.30**	**0.29**	0.24	0.24	0.25	0.29	0.29	**0.29**	-0.02	-0.02	0.00	0.29	0.29	**0.29**	0.28	0.29	**0.29**
Two Norm	**0.34**	**0.30**	**0.30**	0.24	0.18	0.21	0.32	0.29	0.29	0.07	0.01	0.01	**0.34**	0.25	0.27	0.30	0.29	0.29
Pen Based	0.54	**0.27**	**0.28**	**0.59**	0.02	0.00	0.49	0.22	0.22	0.47	0.00	-0.01	0.45	0.17	0.20	0.27	**0.27**	**0.28**
Nursery	**0.30**	**0.30**	**0.28**	0.09	0.18	0.13	0.25	0.24	0.25	0.00	0.01	0.00	0.29	0.29	**0.28**	0.23	0.23	0.22
Magic04	**0.47**	**0.26**	**0.22**	0.42	0.13	0.06	0.41	0.22	0.18	0.32	0.07	-0.06	0.45	0.20	0.13	0.28	0.25	**0.22**

Table 2: RE* metric of FGDI and four state-of-the-art imputation methods for the 13 datasets. Each entry represents the number of times that given algorithm scored RE* < 1 (good estimator) on a total of 138 used features. The median imputation is not reported since it always scores RE* = 1

Dataset (# features)	**FGDI**	**KNNI**	**BTI**	**MICE**	**bPCA**
Contraceptive (9)	9	3	9	1	8
Yeast (6)	5	1	6	0	4
Red Wine (11)	10	6	11	4	9
Car (6)	6	0	5	0	0
Titanic (3)	3	1	3	0	3
Abalone (8)	8	7	8	7	8
White Wine (11)	10	7	10	1	8
Page Block (10)	10	10	10	4	9
Ring (20)	16	0	20	0	14
Two Norm (20)	20	0	20	0	20
Pen Based (16)	15	16	16	14	16
Nursery (8)	8	0	2	0	0
Magic04 (10)	9	8	9	5	9
Total (138)	129	59	129	36	108

As evidenced in Table 3 (first three columns): the imputation improvement achieved by FGDI is statistically significant (*p-value* < *0.05*) in 40 out of the 52 cases (77%); comparable in 9 cases; and worse in 3 cases only. As suggested in [20], the same *NULL* hypothesis was tested using the MAE metric. The FGDI resulted significantly better in 37 cases (71%), comparable in 12 and worse in only 3 cases. The second-best imputation method (bPCA) for RMSE was significantly better in 31 out of the 52 cases (60%); comparable in 9; and worse in 12 cases, which shows an improvement for FGDI of 17% over the best single method. For the MAE hypothesis, bPCA results were significantly better in 24 out of the 52 cases (46%); comparable in 14; and worse in 14 cases, showing inferior imputation accuracy in 25% of the cases, compared with the FGDI. Furthermore, Table 3 shows the robustness of FGDI when estimating the missing values - lower variance than KNNI, MICE, bPCA, and comparable RE* values with BTI (Table 2).

Table 3: RMSE (MAE) significance test for 5-fold cross validation with 25% MCAR, MAR, and MNAR. Each row shows how many times model M_i is better (win), comparable (tie), or worse (loss) than the other models with the Wilcoxon Signed Rank Test

	MCAR			MAR			MNAR		
	win	tie	loss	win	tie	loss	win	tie	loss
FGDI	**40** (**37**)	9 (12)	3 (3)	**41** (**47**)	10 (5)	1 (0)	**47** (**48**)	5 (4)	0 (0)
bPCA	31 (24)	9 (14)	12 (14)	36 (31)	8 (6)	8 (15)	34 (31)	7 (7)	11 (14)
BTI	26 (19)	12 (15)	14 (18)	28 (22)	6 (11)	18 (19)	23 (22)	6 (8)	23 (22)
KNNI	15 (19)	3 (11)	34 (22)	11 (13)	2 (6)	39 (33)	14 (13)	3 (6)	35 (33)
MICE	3 (4)	5 (8)	44 (40)	1 (1)	2 (5)	49 (46)	1 (1)	1 (5)	50 (46)

The following experiments are considered when the missingness is caused by MAR and MNAR mechanisms.

The *Standard Accuracy* values given in Table 1 for the MAR experiment show slightly superior performance of FGDI when compared with the other imputation techniques. The proposed model outperforms the baseline *Random Guessing* ($SA_{FGDI} > 0$) in all reported cases and the *Median Imputation* ($SA_{FGDI} > SA_{Median}$) in 8 out of 13 datasets. Furthermore, it also shows better accuracy in all 13 cases when compared to KNNI and MICE, and superior than BTI and bPCA results in 11 and 10 cases respectively. It is also worth to notice that the imputation under MAR condition is generally harder task (compared to MCAR), since the missingness is not uniformly distributed across the dataset and depends on the other variables as well (as discussed in Section 4). As for all previous experiments, the *Wilcoxon* test is adopted to evaluate the significance in difference for RMSE and MAE metrics. Results in Table 3 (4^{th} to 6^{th} column) show significant imputation improvement of the FGDI for 41 out of the 52 cases (79%); comparable in 10; and worse in only 1 case, when using RMSE. On the other hand, for the MAE metric, the FGDI resulted better in 47 cases (90%); comparable in 5; and never worse. The second-best imputation method (BTI) for RMSE and MAE is significantly better in 36 and 31 out of the 52 cases (69% and 60%); comparable in 8 and 6 cases; and worse in 8 and 15 cases, showing inferior to the FGDI performance in 10% and 30% of the cases respectively.

The same analysis performed under the MNAR condition also suggests that the use of a single imputation method for the whole dataset is not the best option. Again, the SA values (Table 1) are generally lower when compared to the MCAR mechanism as the missingness is caused by the considered variable itself (as explained in Section 4), increasing the likelihood of introducing bias when imputing the values. In the MNAR case, Table 1 also shows superior results for our method in 10 out of 13 datasets. The reported SA_{FGDI} is better than SA_{KNNI} and SA_{MICE} for all considered datasets, while being never worse than SA_{BTI} and SA_{bPCA}. When compared to the baselines, the FGDI is always superior to the *Random Guess* ($SA_{FGDI} > 0$), better than the *Median Imputation* in 7 out of 13 cases, and worse only in 1 of the cases. The Wilcoxon analysis Table 3 (columns 7 to 9) shows the FGDI being better than the second best method (BTI) in 25% and 33% of the cases for RMSE and MAE respectively. Comparing the proposed method with the other imputation techniques, the FGDI is better than bPCA, KNNI and MICE in 46%, 64% and 89% of the cases for the RMSE and 50%, 67% and 90% for the MAE metrics. Despite being generally not recommended [6], the *Median Imputation* showed comparable and even better results than the bPCA, BTI, KNNI, and MICE in both MAR and MNAR settings. At first sight, this result is contradictory to the MCAR experiment (Table 1). This could be explained by the fact that the multivariate model can benefit from the uniformly distributed missingness across the dataset (like in the MCAR mechanism), while for the MAR and MNAR (where the missingness depends on a single variable), the use of a univariate model (baselines) could be reducing the noise in the prediction (because of not considering uncorrelated features). However, as it can be seen from the carried experiments, the use of combination of baselines and state-of-the-art techniques (as in our approach) can improve the accuracy in almost all proposed scenarios with a very low risk of worsening the imputation.

Last point to note is that while the FGDI is superior in all setups, the bPCA and BTI are competing for the second position in the three scenarios (bPCA for MCAR; and BTI for MAR and MNAR). All the experiments presented in this work have been done on a 16-core machine with 32gb RAM and 64Gb SSD of storage. Figure 1 shows the training time for the four state-of-the-art techniques (KNNI, BTI, MICE, and bPCA) and the proposed FGDI method over the 13 datasets, given in seconds. Due to the FGDI parallelization (each imputation algorithm can be run independently from the others), its training execution time is never significantly higher than the time needed for any other single technique. FGDI training time (blue bar in Figure 1) is always comparable with the slowest technique, plus an overhead due to the different scheduled threads. Furthermore, the proposed method shows a consistent time execution overhead with datasets of different volume and features size. This behavior can be observed from the percentage change between the FGDI and the slowest compared model. he percentage change results are smaller for bigger datasets (7.69, 6.15, 14.37, 11.76, 5.84, 10.74 and 4.76 for *White Wine, Page Block, Ring, Two Norm, Pen Based, Nursery* and *Magic04* respectively) and larger for the small ones (22.5, 20, 43.90, 59.09, 56.25, 46.34 for *Contraceptive, Yeast, Red Wine, Car, Titanic* and *Abalone* respectively).

This finding supports the recommendation of using the FGDI regardless the size of the dataset (as long as the imputation is feasible for the single models employed in the FGDI). For the prediction run-time (applied on the test set), FGDI showed to be comparable with the slowest method selected during the training phase.

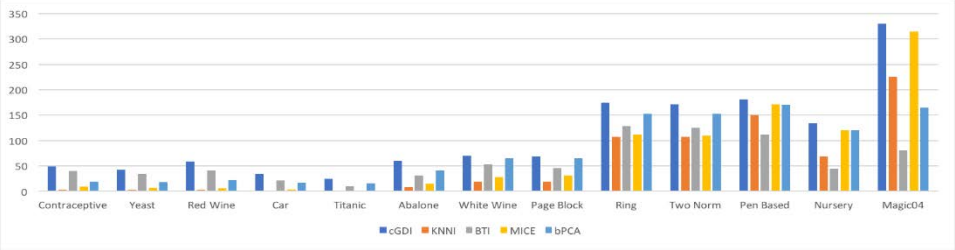

Figure 1: Training time in seconds (y-axis) of the five considered imputation methods over the 13 datasets (x-axis). The Median Imputation is omitted having always a training time less than 1 second

6. CONCLUSION

The investigated FGDI method initially extracts the complete subset (without missing values), and selects through a learning process the most suitable imputation method for each feature. The FGDI imputation performance is evaluated with four widely used metrics for such tasks (SA, RE*, RMSE, and MAE). The results are statistically assessed using the *Shapiro Test* to check the distribution normality, and the non-parametric *Wilcoxon Signed Rank Test*, for statistical significance, using confidence level α=0.05.

Under the MCAR mechanism, the *Standard Accuracy* analysis demonstrates that the proposed model is always more accurate than the baselines and produces better estimation than the state-of-the-art methods in 41 out of 52 cases. The *Wilcoxon* shows improvements of 17% and 25% for the FGDI over the second best performing algorithm (bPCA) over the

two metrics. In addition, FGDI and BTI impute values with higher stability (RE* < 1) for 129 out of 138 tested features, followed by bPCA with 108 out of 138.

Although the prediction under MAR and MNAR mechanisms is generally less accurate than the one under MCAR, the FGDI still shows better performance when compared with the baselines and the state-of-the-art techniques. In particular, in the MAR case, the FGDI is more accurate than the second best model (BTI) in 10% and 30% of the cases for RMSE and MAE respectively. Under the MNAR mechanism the proposed model is again better than BTI in 25% and 33% respectively.

Finally, the performed imputation run time analysis proves the approach feasibility regarding the needed training and testing time. The reported results strongly support the efficiency of the proposed method when implementing multivariate imputation as a way of dealing with missingness. Another advantage is that the FGDI can be easily parallelized, having straightforward implementation allowing other imputation methods to be easily incorporated.

REFERENCES

[1] Enders C. K. (2010) Applied missing data analysis, *Guidford: Guildford Press.*

[2] Schmitt P., Mandel J. and Guedj M. (2015). A comparison of six methods for missing data imputation. *J. of Biometrics & Biostatistics*, 6(1), 1-6.

[3] Jordanov I., Petrov N. and Petrozziello A. (2018). Classifiers Accuracy Improvement Based On Missing Data Imputation. *Journal of Artificial Intelligence and Soft Computing Research*, 8(1), 33-48.

[4] Cohen J., Cohen P., West S. G. and Aiken L. S. (2013). Applied multiple regression/correlation analysis for the behavioral sciences. *Routledge*.

[5] Sarro F., Petrozziello A. and Harman M. (2016). Multi-Objective Software Effort Estimation. *in Software Engineering (ICSE), 2016 IEEE/ACM 38th IEEE Int. Conf. on*, Austin.

[6] Osborne J. and Overbay A. (2008). Best practices in data cleaning. *Best Practices in Quantitative Methods*, 1(1), 205-213

[7] Rahman G. and Islam Z. (2011). A decision tree-based missing value imputation technique for data pre-processing. *in Proc. of the 9th Australasian Data Mining Conf.*

[8] Frènay B. and Verleysen M. (2014). Classification in the presence of label noise: a survey. *Neural Networks and Learning Systems, IEEE Trans. on*, 25(5), 845-869.

[9] Valdiviezo C., and Van Aelst S. (2015). Tree-based prediction on incomplete data using imputation or surrogate decisions. *Information Sciences*, 311, 163-181.

[10] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D. and Altman R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, 17(6), 520-525.

[11] Cartwright M., Shepperd M. J. and Song Q. (2003). Dealing with missing software project data. in *Software Metrics Symposium, 2003. Proc. 9th Int.*

[12] Batista G. and Monard M. (2001). A study of K-nearest neighbour as a model-based method to treat missing data. in *Argentine Symposium on Artificial Intelligence*.

[13] Lee M. C. and Mitra R. (2016). Multiply imputing missing values in data sets with mixed measurement scales using a sequence of generalised linear models. *Computational Statistics & Data Analysis*, 95(1), 24-38.

[14] Graham J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60, 549-576.

[15] Bartlett J., Seaman S., White I. and Carpenter J. (2015)."Multiple imputation of covariates by fully conditional specification: accommodating the substantive model. *Statistical methods in medical research*, 24(4), 462-487.

[16] Oba S., Sato M.-a., Takemasa I., Monden M., Matsubara K.-i. and Ishii S. (2003). A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16), 2088-2096.

[17] Petrozziello A. and Jordanov I. (2017). Column-wise Guided Data Imputation. *Procedia Computer Science*, 108(1), 2282-2286.

[18] Alcalá-Fdez J., Fernandez A., Luengo J., Derrac J., García S., Sánchez L. and Herrera F. (2011). KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3), 255-287.

[19] Pan X.-Y., Tian Y., Huang Y. and Shen H.-B. (2011). Towards better accuracy for missing value estimation of epistatic miniarray profiling data by a novel ensemble approach. *Genomics*, 97(5), 257-264

[20] Willmott C. J. and Matsuura K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, 30(1), 79-82.

[21] Chai T. and Draxler R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?--Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247-1250.

[22] Whigham P. A., Owen C. A. and Macdonell S. G. (2015). A baseline model for software effort estimation. *ACM Trans. on Software Engineering and Methodology (TOSEM)*, 24(3), 20.

[23] Gòmez-Carracedo M., Andrade J., Lòpez-Mahìa P., Muniategui S. and Prada D. (2014). A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. *Chemometrics and Intelligent Laboratory Systems*, 134(1), 23-33.