International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A hybrid modified deep learning data imputation method for numeric data sets

Nuran Peker*[1], Cemalettin Kubat[2]

Submitted: 19/10/2020 Accepted : 02/03/2021

Abstract: Missing data is a major problem in terms of both machine learning and data mining methods. Like most of these methods do not work with missing data, negative results may occur on the performance of the working ones, also. Imputation is a data preprocessing method used to replace missing data with appropriate values. This study aims at developing a hybrid modified imputation method based on deep learning approach. For this purpose, we use Random Forest and Datawig deep learning imputation (called RF-DLI) methods together. Datawig is a deep learning-based library that supports missing value imputation for all types of data. RF-DLI approach includes the following steps to impute missing data. First, the importance of each attribute of the data set is determined with Random Forest (RF). Second, the most important 50% of the attributes are selected. Finally, each missing value is imputed with datawig (DLI) using these most important attributes. The study uses six real-world data sets from different fields with 30% missing data. The imputation performance of RF-DLI is compared to K-Nearest Neighbors (KNN), Multiple Imputation by Chained Equations (MICE), MEAN imputation, and Principle Component Analysis (PCA) imputation approaches in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-square (R^2) evaluation metrics. The results show that in most cases, the RF-DLI approach has better imputation performance than the other techniques mentioned.

Keywords: *Missing values, data imputation, deep learning, random forest*

1. Introduction

Missing data is an important problem frequently encountered by data analysts working with real-world data. The methods have generally been developed under the assumption that the data sets are complete [1,2], but most real-world data sets contain missing values. Many factors play a role in the occurrence of missing data for real-world data sets. For example, absence of response in scientific experiments, power system failures, incorrect data transfer in digital systems, procedure of manual data entry, equipment failure, and environmental factors can lead data breakdown during data collecting and storage processes [3-6]. Analyzes made without taking into account missing data may produce misleading results. In this respect, it is a requirement for researchers to examine the missing data and to take the necessary measures to eliminate the problems that may be encountered before analysis to be carried out on the data.

Missing data are categorized into three missingness mechanism [7]. Missing completely at random (MCAR): missing data that are not related to the variable itself and do not emerge with the effect of another variable and completely randomly distributed within the data set; missing at random (MAR): missing data in the variable are related to other variables but not the change in the variable itself; data are missing not at random (MNAR): missing data in the variable are both related to other variables and to the values of the variable itself. The experiments in this study are conducted under MCAR assumption.

The three most prevalent approaches to overcome the missing data are ignoring records with missing values, single imputation (replacing missing values with median, mean, most frequent, or constant value), and multiple imputation using iterative model-based methods. In cases where missing values are less, it may be appropriate to exclude these values from the data set, but in cases where there are a lot of missing data, this approach causes valuable information to be lost and the analysis process is affected incorrectly [8]. The main drawback of simple imputation is that it disregards the relations between the attributes. Furthermore, this approach does not take into account the distribution of data because it imputes all the missing values for an attribute with the same value. Due to the inadequacy of the aforementioned approaches, a number of other approaches also including multiple imputation are proposed. A widespread multiple imputation method is iterative principal component analysis [9] that used to imputation flow network [10]. Multivariate Imputation by Chained Equations (MICE) [11] is another multiple imputation-based approach. In the method, missing values are replaced with reasonable values to predict more truthful regression coefficients that are not affected by missing data. A regression model of decision trees [12] has been successfully used for multiple imputation of industrial databases. There are other machine learning-based imputation techniques such as K-Nearest Neighbors (KNN) [13], Kohonen Self-Organizing Map (SOM) [14], Probabilistic Neural Networks (PNN) [15], and clustering-based approaches [16,17].

In this study, we combine RF and DLI methods to impute missing data. RF is a machine learning approach used for purposes such as classification, regression, feature extraction. DLI is a deep learning-based data imputation method. However, DLI is used to impute missing data just in a single column of a data set. In our study, in order to eliminate this constraint of DLI, we extract the

[1] *Industrial Eng., Sakarya University, Sakarya-54050, TURKEY ORCID ID:0000-0003-3040-3962*

[2] *Industrial Eng., Sakarya University, Sakarya-54050, TURKEY ORCID ID:0000-0002-8666-1517*

* *Corresponding Author Email: nuranpeker41@gmail.com*

most important attributes of each data set with RF and then complete the missing values in all columns of the data set with DLI, using these important attributes we obtained.

2. Related Works

Today, many deep learning-based imputations approaches are suggested by researchers. The study [18] recommends an approach for traffic data imputation based on deep learning. In the study, the deep learning approach finds the correlations between data structures to improve the imputation accuracy. DAPL[19] is an alternative deep learning-based imputation method for gene expression and DNA data. SAIC [20] introduces a new hybrid method of missing value imputation that combines stacked auto-encoder and incremental clustering. Experimental results show that the method effectively fills in the missing data as well as improves the time performance. [21] is another hybrid missing data imputation method that uses neural networks and weighted KNN together. The study [22] proposes a hybrid approach that uses a neural network and genetic algorithm together to impute the missing values for medical IoT implementations. The method benefits from deep learning for predicting the missing data and genetic algorithm for optimizing the weights of the neural network. The study [23] merges deep neural networks, genetic algorithms, maximum likelihood estimator, and swarm intelligence to impute monotone and arbitrary missing data. Method [24] proposes an imputation model using the autoencoder-based architecture that reduces the complexity of the data. MIDA [25] is a model based on overcomplete deep denoising autoencoders that can deal with different data types, missingness proportions, and missingness patterns. The study [26] offers a probabilistic imputation approach based on deep generative models that can get nonlinear relationships among observed values and missing entries for missing data. Work [27] defines a new deep learning neural network imputation approach and its implementation to impute assay bioactivity values. The methods developed in the studies mentioned are generally tested on a data set in a single field. It is not possible to talk about a method that produces excellent results on all data sets in the literature. In our study, using 6 data sets picked from different fields, we show that our method can be applied successfully in a wider area.

3. Material and Methods

3.1. Data sets

Data sets used in this study are taken from UCI Machine Learning Repository [28]. All of these data sets in various sample sizes consist of continuous variables. The "*id*" attribute has been removed from the Glass, Ecoli, and Yeast data sets. The values of the data sets are shown in Table 1.

Table 1. Data sets used in this study.

Dataset	#Attributes	#Examples	Attribute type	#Classes
Vertebral	6	310	continuous	2
Glass	10	214	continuous	7
Seed	7	210	continuous	3
Ecoli	8	336	continuous	8
Vehicle	18	846	continuous	4
Yeast	9	1484	continuous	10

3.2. Imputation Methods

In this study, we compare the performance of the RF-DLI with four methods well known and commonly used in the literature.

KNNimputation: KNN is actually a classification method. In KNNimputation [13] this classification logic is used to fill missing values. KNNimputation is based on the proximity of the observations to each other. In the method, missing values are imputed with the average of 'K' nearest observations by measuring the distance between variables that do not contain missing values. The chosen "K" value is quite important for the performance of the algorithm. Because if K is lower than necessary causes to low bias and high variance, higher than necessary causes to high bias and low variance. Thus the selection of K should be made carefully.

MICEimputation: MICE [11] approach imputes each missing data point multiple times, considering the data distribution of the observation values. In this way, it is possible to take into account the uncertainty around the observation values and to make more unbiased estimates. Thus, multiple completed data sets are obtained. Then, these data sets are analyzed and the results are combined. Since the method has also a very flexible structure, it can be carried out to different data types.

MEANimputation: In this method [29], the mean value of each variable is calculated ignoring missing values. Then, each missing value is filled with this mean value of the variable it belongs to. The method is relatively easy to implement but it does not provide well performance for all situations. The method does not take into account the relationship between the data. Since it always fills in missing data with the same value, it can cause deviations in the variance of the data and standard errors. This situation negatively affects the statistical analysis.

PCAimputation: In this method, missing values of mixed data are imputed using the regularised iterative PCA algorithm [30]. The algorithm first imputes missing data with initial values such as the mean of the variable. In the second step regularized iterative PCA algorithm performs PCA [31] on the completed data set. Then, it imputes the missing values with the reconstruction formula of order the number of components (the fitted matrix computed with components for the regularized scores and loadings). These steps of prediction of the parameters via PCA and imputation of the missing values using the fitted matrix are iterate until convergence.

3.3. RF-DLI

In this study, we use a hybrid imputation approach that combines RF with modified DLI to impute missing values.

RF: Feature selection is a very important issue in data science. Because high dimensional data sets have high computational complexity and low data interpretability. Random forest [32] is one of the frequently used feature selection algorithms thanks to its ease of application. It has a low overfitting, good classification performance, and easy interpretability, in general. The importance of the features is determined according to the calculated impurity value. This importance varies in parallel with how much the feature reduces impurity. The feature that reduces impurity more is more important. The impurity value in RF is calculated using Gini index [31], information gain [32], or variance. We use variance in our study. The feature importances of data sets determined by RF are shown in Figure 1. The pseudocode of RF-DLI is shown in Figure 2.

DLI: DLI [33] is a deep learning-based software package that imputes the missing values for heterogeneous data types. It tunes hyperparameters of all input columns, except the column to be imputed, automatically. Then it trains a classifier on these features to predict the missing data. It performs domain adaption by determining and correcting the difference between rows that are imputed and rows that are used for training. In the process of DLI, for each column to be imputed (output column), the user has to specify the columns (input columns) which can contain helpful information for imputation.

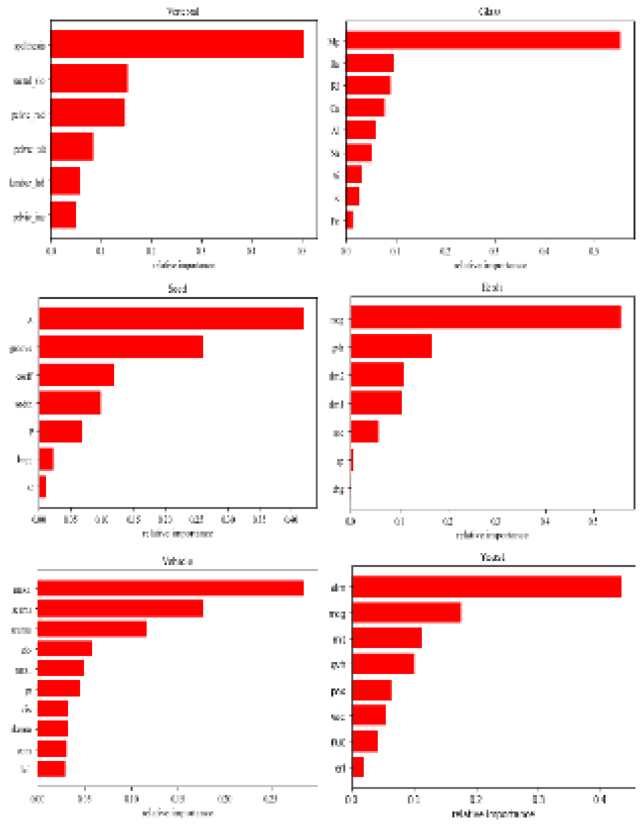

Fig. 1: Feature importances of data sets

```
Input: Dataset D with no missing values
Step 1: Determine importance of attribute of D with RF
Step 2: Select x, that the most important %50 of attributes of D
        determined by RF
Step 3: Split D randomly into dataTrain(complete)
        and dataTest(incomplete)
Step 4: Perform following:
      4.1: input_columns: columns containing information about
           the column to be imputed
      4.2: output_column: the column to be imputed
Step 5: Create a list(predictedList[]) to add predictions
Step 6: for i in columns D
        6.1: input_columns  = x
        6.2: output_column  = i
        6.3: Train DLI with dataTrain
        6.4: Predict missing values of i in dataTest with trained DLI
        6.5: Append i_predicted to predictedList[]
        end for
Step 7: return predictedList[]
Output: Complete data
```

Fig. 2. Pseudocode of RF-DLI routine.

3.4. Evaluation Metrics

In this study, the performance of the imputation methods used are compared using mean absolute error (MAE) [34], root mean square error (RMSE) [35], and R-square (R^2) [36].

MAE: It is the measure of the absolute difference between two continuous variables. In other words, MAE is a linear score that measures the average magnitude of errors in a series of estimates without considering their direction, in which all singular errors are weighted equally on the mean. MAE can range from 0 to ∞. MAE with low value denotes better model performance. It is calculated as in Equation 1.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (1)$$

RMSE: It is a quadratic metric that is used to find the distance between the values predicted by a model and the actual values of the data. RMSE is the standard deviation of the prediction errors. RMSE can range from 0 to ∞. The lower RMSE value, the higher the performance of the model. It is calculated as in Equation 2.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (2)$$

R^2: R^2 (the coefficient of determination) is the ratio of the variance in the dependent variable that can be estimated from the independent variables. R^2 normally takes values between 0 and 1, but negative results may also occur due to the difference in the computation method. R^2 values close to 1 mean that the model performance is good. The calculation method we use in our study can produce negative R^2 values, but no negative values are encountered in our experiments. It is calculated as in Equation 3.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \qquad (3)$$

In Equation 1, Equation 2, and Equation 3, y_i is observed value, $\hat{y}_i$ is predicted value of y_i , and $\bar{y}$ is mean value of y.

4. Experimental Results

In this study, we use six data sets that selected from different fields. We divide complete data sets into training data (%70) and testing data (%30) randomly

Table 2. Comparison of imputation methods (missing rate =30%)

Dataset	Metric	KNN	MICE	MEAN	PCA	RF-DLI
Ecoli	MAE	0.0335	0.0327	0.0335	0.0292	**0.0175**
	RMSE	0.0916	0.0974	0.0883	0.0755	**0.0473**
	R^2	0.5360	0.6721	0.5545	**0.7783**	0.5807
Vehicle	MAE	4.9859	6.7878	5.4045	5.7048	**3.5828**
	RMSE	25.339	31.86	24.485	25.156	**7.7921**
	R^2	0.6637	0.4161	0.6942	0.6772	**0.8259**
Vertebral	MAE	3.5311	4.2848	4.2115	**3.1240**	3.8557
	RMSE	12.529	12.028	13.029	11.611	**6.9668**
	R^2	0.7289	0.6094	0.6584	0.7534	**0.7718**
Yeast	MAE	0.0208	0.0221	**0.0193**	0.0184	0.0203
	RMSE	0.0616	0.0713	0.0572	**0.0543**	0.0563
	R^2	0.6201	0.4811	0.6698	**0.6920**	0.4856
Glass	MAE	**0.1124**	0.1318	0.1347	0.1159	0.1274
	RMSE	0.3993	0.4731	0.4277	0.3807	**0.2477**
	R^2	**0.7576**	0.6961	0.7233	0.7443	0.6460
Seed	MAE	0.1288	0.0801	0.2589	0.0896	**0.0366**
	RMSE	0.4438	0.3554	0.7448	0.3446	**0.0592**
	R^2	0.8876	0.9256	0.7152	0.9120	**0.9808**

In our model assumption, all values of testing data are missing. Because the model returns a predicted column for each column of testing data. In the training process, the imputation model is trained with default hyperparameters on training data. In the imputation process, the true values in the testing data columns are estimated using the predictions obtained with training data. More formally, RF-DLI imputes values y_{out} = f(x_{inp}) where f states to the imputation model trained on the actual values in the column out, x_{inp} states the most important 50% feature extracted by RF, and y_{ou} states the column to be imputed. We compare RF-DLI on numerical data sets containing 30% missing data by four different methods: KNN, MICE, MEAN, and PCA. Scikit learn [37] machine learning library is used in the implementation of the KNN and MEAN. CRAN-R [38] software is selected as the main tool for the implementation of MICE and PCA. In KNN, the value of k is taken as 5. In MICE, multiple iterations are taken as 5, the imputation method is set to the predictive mean matching (for numeric values), and the number of iterations to 50. The number of components is set to 2 in PCA. All of these values are default parameters for the mentioned methods. For the data sets that imputed, MAE, RMSE, and R^2 values have been given in Table 2. For the data sets that imputed with the RF-DLI, model performance plotting is shown in Figure 3. As seen in Table 2, the lowest MAE value is obtained with the RF-DLI algorithm for the Ecoli, Vehicle, and Seed data sets. For all data sets except Yeast, the lowest RMSE value is found with the RF-DLI. The highest R^2 value for the Vehicle, Vertebral, and Seed data sets get with the RF-DLI approach. PCA finds the best R^2 value for the Ecoli and Yeast data sets. Also, PCA is more successful in terms of MAE for the Vertebral data set and RMSE for the Yeast data set. KNN is more successful in terms of MAE and R^2 for the Glass data set. MEAN finds the best result for MAE on the Yeast data set. Based on the previously mentioned definitions of the metrics used, considering the MAE, RMSE, and R^2 values obtained by the Rf-DLI, it is seen that the absolute errors and standard deviation of errors between the predicted values of the model and the actual values of data are at a highly desirable level, especially for Ecoli, Yeast, and Seed data sets. When the values obtained by the model are considered as a whole, it can be said that the model generally produces estimates closer to the real data compared to the other methods. In addition, the method uses 50% attributes that best explain the variance in the data for each data set during the imputation process. This means that the features included in the calculation are reduced and therefore the computational complexity is also decreased.. The mean metric values of the methods in the experiments performed on the data sets are shown in Figure 4. As seen in Figure 4, considering the mean metric values, PCA produces more successful result for R^2. Apart from that, the RF-DLI method is more successful than the other methods.

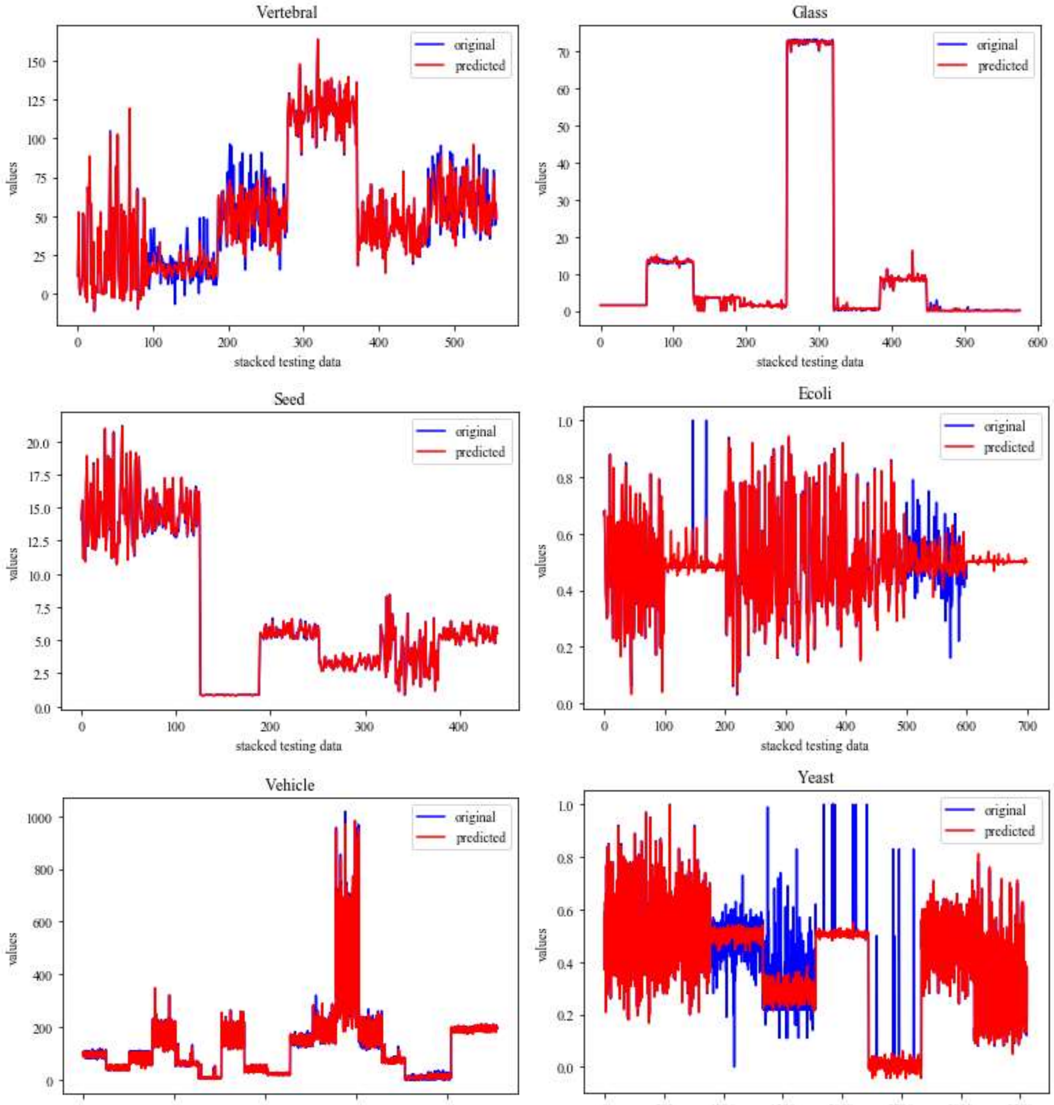

Fig. 3. The performance of the RF-DLI on data sets

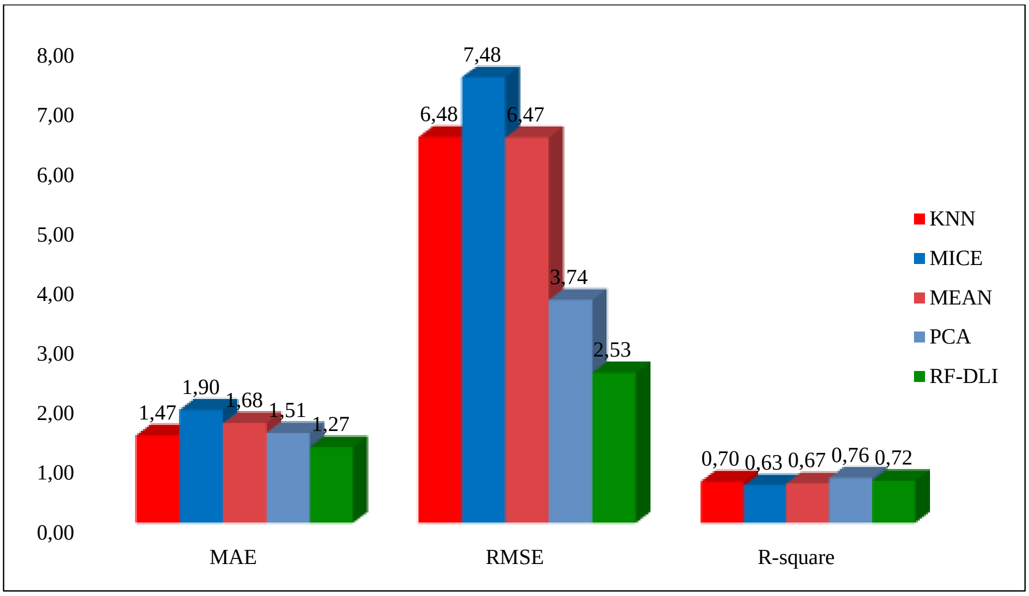

Fig. 4. Mean metric values of the methods

5. Conclusions and Future Works

This study aims to present a hybrid method, RF-DLI, on a deep learning framework that imputes missing data for numeric data sets. In the method, the most important features of the data sets are extracted with RF, and only these features are used as input columns in the imputation process. The RF-DLI method is compared with KNN, MICE, MEAN, and PCA in a series of experiments performed on six real-world data sets. The mean MAE value of KNN, MICE, MEAN, PCA, and RF-DLI is 1.47, 1.90, 1.68, 1.51, and 1.27 for the data sets, respectively. The average RMSE value of KNN, MICE, MEAN, PCA, and RF-DLI is 6.48, 7.48, 6.47, 3.74, and 2.53 for the data sets, respectively. The average R^2 value of KNN, MICE, MEAN, PCA, and RF-DLI is 0.70, 0.63, 0.67, 0.76, and 0.72 for the data sets, respectively. The results acquired show that the RF-DLI method is more successful than the other approaches, generally.

For future works, we will try our method on data sets with mixed types of data. In addition, we aim to test the success of the model on data sets with higher missing values such as 40% and 50%.

References

[1] P. D. Allison, "Missing data techniques for structural equation modeling," Journal of abnormal psychology, 112(4), 545, 2003.

[2] T. D. Pigott, "A review of methods for missing data," Educational research and evaluation, 7(4), 353-383, 2001.

[3] M. Amiri, R. Jensen, "Missing data imputation using fuzzy-rough methods," Neurocomputing, 205, 152-164, 2016.

[4] G. Rahman, Z. Islam, "A decision tree-based missing value imputation technique for data pre-processing," In Proceedings of the Ninth Australasian Data Mining Conference-Volume 121, pp. 41-50, Dec. 2011.

[5] H. Wang, S. Wang, "Mining incomplete survey data through classification," Knowledge and information systems, 24(2), 221-233, 2010.

[6] A. Farhangfar, L. Kurgan, J. Dy, "Impact of imputation of missing values on classification error for discrete data," Pattern Recognition, 41(12), 3692-3705, 2008.

[7] D.R. Rubin, "Inference and missing data," Biometrika, 63(3), 581-592, 1976.

[8] R. J. Little, D. B. Rubin, "Statistical analysis with missing data," Vol. 793, John Wiley & Sons., 2019.

[9] S. Dray, J. Josse, "Principal component analysis with missing values: a comparative survey of methods," Plant Ecology, 216(5), 657-667, 2015.

[10] S. A. Imtiaz, S. L. Shah, S. Narasimhan, "Missing data treatment using iterative PCA and data reconciliation," IFAC Proceedings Volumes, 37(9), 197-202, 2004.

[11] S. V. Buuren, K. Groothuis-Oudshoorn, "mice: Multivariate imputation by chained equations in R," Journal of statistical software, 1-68, 2010.

[12] K. Lakshminarayan, S. A. Harp, T. Samad, "Imputation of missing data in industrial databases," Applied intelligence, 11(3), 259-275, 1999.

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

[14] L. Folguera, J. Zupan, D. Cicerone, J. F. Magallanes, "Self-organizing maps for imputation of missing data in incomplete data matrices," Chemometrics and Intelligent Laboratory Systems, 143, 146-151, 2015.

[15] K. J. Nishanth, V. Ravi, "Probabilistic neural network based categorical data imputation," Neurocomputing, 218, 17-25, 2016.

[16] B. M. Patil, R. C. Joshi, D. Toshniwal, "Missing value imputation based on k-mean clustering with weighted distance," In International Conference on Contemporary Computing, pp. 600-609, Springer, Berlin, Heidelberg, Aug. 2010.

[17] N. Ankaiah, V. Ravi, "A novel soft computing hybrid for data imputation," In Proceedings of the International Conference on Data Science. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2011.

[18] Y. Duan, Y. Lv, W. Kang, Y. Zhao, "A deep learning based approach for traffic data imputation," In 17th International IEEE Conference on Intelligent Transportation Systems (ITSC) , pp. 912-917, IEEE, Oct. 2014.

[19] Y. L. Qiu, H. Zheng, O. Gevaert, "A deep learning framework for imputing missing values in genomic data," bioRxiv, 406066, 2018.

[20] L. Zhao,Z. Chen, Z. Yang, Y. Hu, "A hybrid method for incomplete data imputation," In 12th International Conference on Embedded

Software and Systems, pp. 1725-1730, IEEE, Aug. 2015.

[21] I. B. Aydilek, A. Arslan, "A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks," International Journal of Innovative Computing, Information and Control, 7(8), 4705-4717, 2012.

[22] N. Al-Milli, W. Almobaideen, "Hybrid neural network to impute missing data for IoT applications," In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), pp. 121-125, IEEE, Apr. 2019.

[23] Leke, C., Marwala, T., & Paul, S. "Proposition of a theoretical model for missing data imputation using deep learning and evolutionary algorithms," arXiv preprint arXiv:1512.01362, 2015.

[24] X. Lai, X. Wu, L. Zhang, W. Lu, C. Zhong, "Imputations of missing values using a tracking-removed autoencoder trained with incomplete data," Neurocomputing, 366, 54-65, 2019.

[25] L. Gondara, K. Wang, "Mida: Multiple imputation using denoising autoencoders," In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 260-272, Springer, Cham, Jun. 2018.

[26] H. Zhang, P. Xie, E. Xing, "Missing value imputation based on deep generative models," arXiv preprint arXiv:1808.01684, 2018.

[27] T. M. Whitehead, B. W. J. Irwin, P. Hunt, M. D. Segall, G. J. Conduit, "Imputation of assay bioactivity data using deep learning," Journal of chemical information and modeling, 59(3), 1197-1204, 2019.

[28] A. Asuncion, D. Newman, UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/index.php. Accessed on August 21, 2020

[29] R. J. Little, D. B. Rubin, "Statistical analysis with missing data," John Wiley & Sons, 1987.

[30] Josse, J., Husson, F, "Handling missing values in exploratory multivariate data analysis methods," Journal de la Société Française de Statistique, 153(2), 79-99, 2013.

[31] Hotelling, H, "Analysis of a complex of statistical variables into principal components," Journal of educational psychology, 24(6), 417,1933.

[32] L. Breiman, "Random forests," Machine learning, 45(1), 5-32, 2001.

[33] C. Gini, "Variabilità e mutabilità," Vamu, 1912.

[34] J. R. Quinlan, "Induction of decision trees," Machine learning, 1(1), 81-106, 1986.

[35] F. Biessmann, T. Rukat, P. Schmidt, P. Naidu, S. Schelter, A. Taptunov, D. Salinas, "DataWig: Missing value imputation for tables," Journal of Machine Learning Research, 20(175), 1-6, 2019.

[36] Willott, C. J., & Matsuura, K. "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, 2005.

[37] Barnston, A. G. "Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score," Weather and Forecasting, 7(4), 699-709, 1992.

[38] Barten, A. P. "The coefficient of determination for regression without a constant term.," In The Practice of Econometrics (pp. 181-189). Springer, Dordrecht, 1987.

[39] Pedregosa et al., "Scikit-learn: Machine Learning in Python," JMLR 12, 2825-2830, 2011, Accessed on Sep. 10, 2020

[40] R Foundation for Statistical. [Online]. Availible: https://www.R-project.org, 2016, Accessed on Sep. 10, 2020.