Normalization and Outlier Removal in Class Center-Based Firefly Algorithm for Missing Value Imputation

Heru Nugroho (heru@tass.telkomuniversity.ac.id)
Institut Teknologi Bandung https://orcid.org/0000-0002-7460-7687
Nugraha Priya Utama
Institut Teknologi Bandung
Kridanto Surendro
Institut Teknologi Bandung

Research

Keywords: missing data, normalization, outliers, C3-FA, class center

Posted Date: June 9th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-538193/v1

Normalization and Outlier Removal in Class Center-Based Firefly Algorithm for Missing Value Imputation

Heru Nugroho[1]*, Nugraha Priya Utama[2], Kridanto Surendro[3]

School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia

**Corresponding author: heru@tass.telkomuniversity.ac.id*

Abstract

Missing data is one of the factors often causing incomplete data in research. Data normalization and missing value handling were considered major problems in the data pre-processing stage, while classification algorithms were adopted to handle numerical features. Furthermore, in cases where the observed data contains outliers, the missing values' estimated results are sometimes unreliable, or even differ greatly from the true values. This study aims to proposed combination of normalization and outlier removal's before imputing missing values using several methods, mean, random value, regression, multiple imputation, KNN, and C3-FA. Experimental results on the sonar dataset show normalization and outlier removal's effect in these imputation methods. In the proposed C3-FA method, this produced accuracy, F1-Score, Precision, and Recall values of 0.906, 0.906, 0.908, and 0.906, respectively. Based on the KNN classifier evaluation results, this value outperformed the other five (5) methods. Meanwhile, the results for RMSE, Dks, and r obtained from combining normalization and outlier removal's in the C3-FA method were 0.02, 0.04, and 0.935, respectively. This shows that the proposed method is able to reproduce the real values of the data or the prediction accuracy and maintain the distribution of the values or the distribution accuracy.

Keyword: missing data, normalization, outliers, C3-FA, class center

1. Introduction

When people consider the heterogeneity of certain sources, the specificity of the source becomes clearer. For example, data streams from a sensor network can be characterized in terms of their quality by the fact that data are often missing and, if they are not missing, are potentially exposed to significant noise and calibration effects[1]. In most research, missing value is a common and serious problem, often leading to biased, inaccurate, and unreasonable conclusions, in cases of inappropriate handling [2–11]. Currently, available analytical methods only have the capacity to work with complete data [12–15]. Thus, missing data-related problems present research opportunities to obtain the right technique to serve as a solution [16].

In classification problems, missing values is a general weakness with the capacity to produce ineffective prediction system results [12,17,18]. Therefore, ignoring missing data affects analysis results [2,9,19–21], learning outcomes, as well as prediction results on collaborative prediction problems [22], and even has the potential to weaken results and conclusion validities [4,21]. In the predictive model, incorrect selection of the missing data handling method tends to affect the model's performance [9,23] as well as the classifiers' accuracy and performance [24]. Previous studies have produced a class center-based adaptive approach model for imputing missing data [25], as the development of methods by considering correlation [26]. However, these studies failed to consider data normalization and outlier detection before performing the imputation process.

Previous research results show that feature normalization has an important impact on classification accuracy. [27–30]. In a dataset with numeric feature attributes, data normalization and processing of missing values are regarded as the main problems in the preprocessing stage. [31]. The normalized mean interpolation method is developed to solve the missing data value in numerical data sets [32]. Numerous studies have separately analyzed the effects of various normalization techniques and strategies for dealing with missing value on classification performance. However, only a few study have rated the effects combining the two [31]. Applying data normalization has a significant impact on classification performance and greatly improves the performance of the KNN imputation method [33]. Previous studies have also shown combining normalization and imputation using the mean, produces more accurate than traditional mean and median methods. [32].

The model-driven imputation algorithm requires that the observable data has no missing values in the dataset, so the characteristics of the observable data directly affect the results of the imputation [34]. Training data usually contains noisy data or outliers that will affect the final performance of the trained model [35,36]. From an instance selection perspective, the dataset is bound to contain some noisy data or outliers, being observed for missing value imputations. Thus, instance selection is conducted to filter out some noisy data, as well as unrepresentative outliers from a given (training) dataset, and the selection's performance must be assessed before imputing missing values [34]. Meanwhile, in the linear regression method, ordinary least squares (OLS) are used to estimate the model's parameters. However, the presence of outliers makes the estimation of these parameters unreliable [37]. Thus, the imputation result is not good enough to fulfill the given precision [38], and has a negative effect on the values entered for missing data [39]. Outlier handling must therefore be performed before imputation, to solve this problem [38,39]. The classical method is unable to accurately conduct imputation in the presence of outliers [40], thus, researchers suggest several imputation methods to overcome these problems [39,41–43].

This study aims proposed combination of normalization and outlier removal's effects on several imputation methods, and compare the results with the class center-based adaptive approach model for missing data imputation conducted in the previous research. Furthermore, this study's novelty is an evaluation of normalization and outlier detection combination's impact on several missing data imputation methods. Also, no previous studies have used this combination simultaneously in an adaptive model of missing data imputation, based on the class center. The combination of outlier detection and normalization in the Firefly Algorithm for handling missing data based on class center is an efficient technique to obtain the data's true value, and maintain the true data values' distribution.

In this study, part II discusses analysis relating to missing data, including methods for handling missing data, normalization and outlier detection's effects on imputation methods, including the FA algorithm to be used in the imputation process and evaluation model applied. In addition, part III contains experimental results suggestions discussing the stages of the research being conducted, while part IV contains the discussion and conclusions.

2. Related Work

A. Methods for Handling Missing Data

The method of dealing with missing data largely depends on the data type and requirements. There are two imputation methods: statistics and machine learning [44,45]. The statistical methods widely used in previous studies are Expectation maximization (EM), Linear regression (LR), Least squares (LS), and mean/mode.

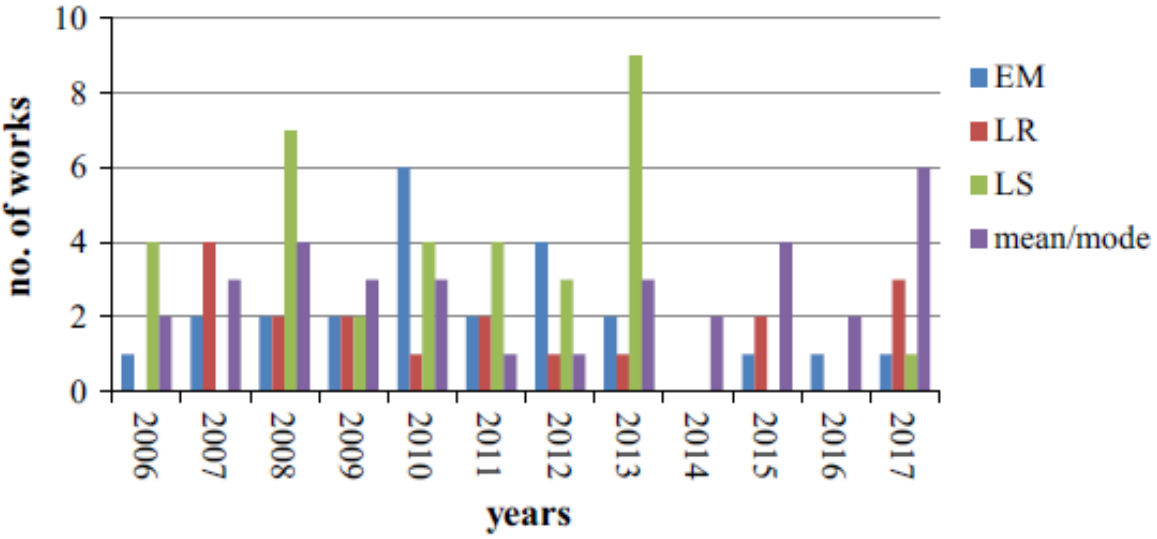


Fig 1. Distribution of the number of studies using EM, LR, LS, and mean/mode techniques [46].

Meanwhile, the commonly used machine learning methods are Decision tree (DT), clustering, K-nearest neighbor (KNN), and Random forest (RF) [46].

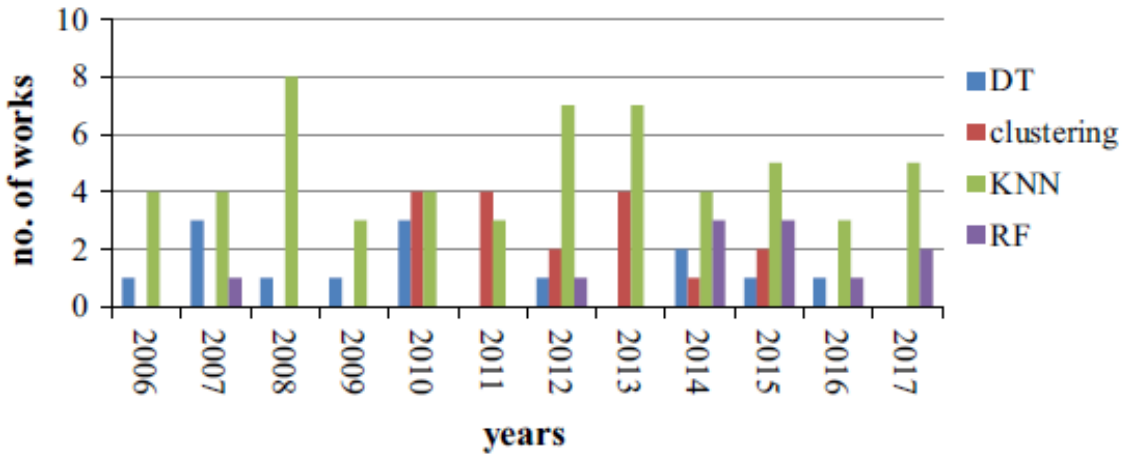


Fig 2. Distribution of the number of studies using DT, clustering, KNN, and RF techniques [46].

Among the various machine learning algorithms, the KNN algorithm is widely used to suggest missing data due to its simple implementation and relatively high accuracy [15,44,47–49].

B. Normalization and Outlier Detection in Missing Data Imputation Algorithm

Three mandatory technical issues that must be considered in the process of inputting missing data, selecting experimental data sets, methods used, and evaluating imputation results is shows in Figure 3 [46].

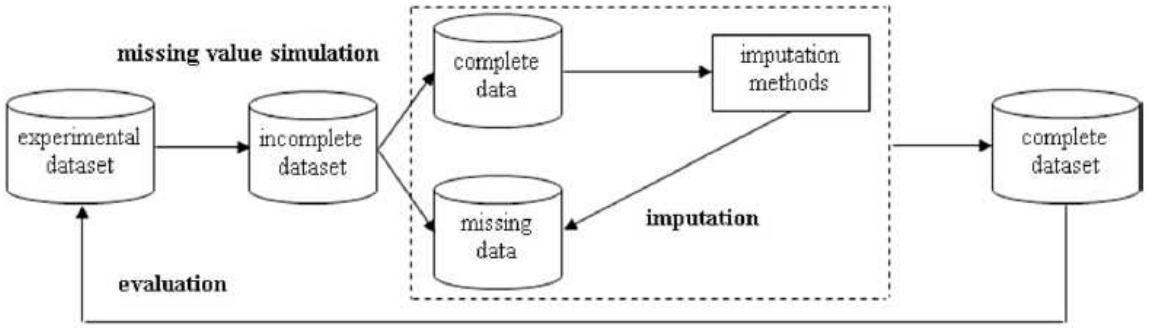


Fig 3. The experimental design procedure for missing data imputation experimental [46].

In addition, the choice of the experimental data set is related to the problem area, the filling of test data, the type of test data, the type of missing data mechanism (MCAR, MAR, MNAR) and percentage (missing rate). According to Lin and Tsai (2020), the normalization and outlier detection's consideration has not been discussed

in the review paper "Missing value imputation: a review and analysis of the literature (2006-2017)". The effect of normalization and/or various techniques for handling the missing value strategy on classification performance separately, has been extensively conducted in previous research. However, only a few studies have assessed the effects of the simultaneous combination of standardization and missing data handling methods (Alshdaifat et al., 2021). Previous studies have also shown combining normalization and imputation techniques produces better accuracy values [32,33,50].

In addition to normalization in pre-processing, outliers significantly influence the statistical estimation process (for instance, the sample mean and standard deviation), resulting in either excessively high or excessively low values [51]. Several missing data imputation methods including mean, linear regression, multiple, and class center-based, utilize the mean value. Generally, the training data contains noisy data or outliers with the ability to reduce the learning model's final performance [35,36]. Therefore, it is necessary to select instances in the observed data set for imputation of missing values and to determine the performance of the selection of instances from the observed data set prior to that imputation [34]. Meanwhile, other studies state outliers play an important role in the imputation method's performance. In cases where a dataset contains outliers, mixed models with high flexibility are able to produce deviations from the true data pattern [52].

Also, previous studies reported imputation results to be strongly influenced by the presence of outliers [37–39]. Therefore, outlier handling must be conducted before imputation [38,39]. Currently, the classical method is unable to perform imputation accurately in the presence of outliers [40], however, various methods have been proposed as a solution to this problem [39,41–43]

C. Class center-based firefly algorithm for handling missing data

The pattern of fireflies with a lower light intensity was used and approximated the group of fireflies with a lower intensity when the missing data was entered. Fireflies with less light are analogous to the missing data attribute, while the intensity counterparts that are brighter are analogous to the complete data attribute. Also, the class center, as the basis of the imputation is used as the objective function $f(x)$, and therefore serves as the prefix in determining the value of $I(x)$.

The pseudocode below summarizes the Firefly Algorithm's main steps for handling missing value based on class center [25].

1. Incomplete data sets are divided into complete and incomplete subsets.
2. Calculate the class center, ($centD_i$) and standard deviation (std) for each class i of the complete subset.
3. Calculate the distance between class center $centD_i$ and other data samples in class i, using the Euclidean distance.

$$Dis(cent(D_i), j) = \sqrt{\left(x_i - cent(D_i)\right)^2} \qquad (1)$$

4. Compute attribute correlations (R) for the complete subset.

$$R_{x_1 x_2} = \frac{n\sum x_1 x_2 - \left(\sum x_1\right)\left(\sum x_2\right)}{\sqrt{\left(n\sum {x_1}^2 - \left(\sum x_1\right)^2\right)\left(n\sum {x_2}^2 - \left(\sum x_2\right)^2\right)}} \qquad (2)$$

5. For each attribute in the incomplete dataset, a value (x) is calculated based on the objective function f(x) and class center values.

$$I(x) = \frac{1}{cent(D_i)} \qquad (3)$$

6. Find a value $I(x) = \frac{1}{x_i}$ the greatest value $I(x) = \frac{1}{CentD_i}$. In cases where there is data containing the largest I (x), update the data movement $x_{i\_new}^k$ using the following movement equation with the assumptions $\beta_0 = 1$, $r = Dis(cent(D_i), j)$ and $\alpha \in [0,1]$.
   a. The formula below is used in cases where the class center value ($CentD_i$) of the attribute containing missing data is equal to the correlated attribute data's $CentD_i$.

$$x_{i\_new}^k = x_{i\_old}^k + \beta_0 e^{-\gamma r^2}\left|centD_i - x_{i\_old}\right| + \alpha\left(rand - \frac{1}{2}\right) \qquad (4) \text{ with } \gamma = centD_i$$

   b. In cases where the $CentD_i$ of the attribute with missing data is below the correlated attribute data's $CentD_i$, the formula below is used.

$$x_{i\_new}^k = x_{i\_old}^k + \beta_0 e^{-\gamma r^2}\left|centD_i - x_{i\_old}\right| + \alpha\left(rand - \frac{1}{2}\right) \qquad (5) \text{ with } \gamma = \left({centD_i}/{R}\right) + \left|diff\ of\ centD_i\right|$$

   c. The formula below is used in cases where the $CentD_i$ of the attribute containing missing data is greater, compared to the correlated attribute data's counterpart.

$$x_{i\_new}^{k} = x_{i\_old}^{k} + \beta_0 e^{-\gamma r^2} \left| centD_i - x_{i\_old} \right| + \alpha \left(rand - \frac{1}{2} \right) \quad (6) \text{ with}$$

$$\gamma = \left(centD_i \times R \right) - \left| diff\ of\ centD_i \right|$$

7. Analyze the imputation's result by comparing the distance between the data and class center obtained from the previous imputation value + - *stdev*. This result is determined based on the closest distance.

D. Performance evaluation

To observe the effect of normalization and outlier detection on several imputation methods, evaluation of machine learning models, for instance, AUC, Precision, Recall, and F1-Score based on the Confusion Matrix, are used. This matrix is a very popular method for solving classification problems, and is suitable for binary classification problems and classification problems with multiple classes [53].

Table 1. Confusion matrix for binary classification.

		Actual	
		Positive	**Negative**
Predictions	**Positive**	TP	FP
	Negative	FN	TN

The confusion matrix represents the predicted and actual state of the data generated by the machine learning algorithm. Precision is the relationship between true positive prediction and overall positive prediction.

$$\text{Precesion} = \frac{TP}{TP + FP} \quad (7)$$

Meanwhile recall (Sensitivity) is the relationship between the true positive prediction and the overall true positive data.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (8)$$

The F1 Score is a weighted average comparison of precision and recall.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \quad (9)$$

AUC (Area under the Curve) is the ROC (Receiver Operating Characteristic), a curve depicting probability with sensitivity and specificity variables, with a limit value between 0 and 1. This area provides an overview of the

model's overall top measurement suitability, and is a standard measure used to indicate the prediction result's quality [54].

In addition to using machine learning evaluation models, for instance, models based on the Confusion Matrix, the proposed imputation method is to be evaluated based on two factors, Predictive Accuracy (PAC), concerned with the imputation technique's efficiency in obtaining the true data value, and Distributional Accuracy (DAC). For the PAC assessment, two measures, Pearson Correlation Coefficient (r) and Root Mean-Squared Error (RMSE), are used [55,56] .

Pearson's correlation coefficient uses variance to measure the correlation between the imputed and actual values, and the degree to which data points tend to deviate from the mean [57]. An effective imputation method should be close to 1[55,56]. In cases where x is the attribute value in the complete data and $\hat{x}$, the correlation coefficient is calculated according to the formula (10).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)(\hat{x}_i - \bar{\hat{x}}_i)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 \sum_{i=1}^{n}(\hat{x}_i - \bar{\hat{x}}_i)^2}} \qquad (10)$$

The root mean square error (RMSE) is a well-known main criterion used to compare the performance of prediction methods by measuring the difference between the estimated value of a given characteristic and the baseline value. In this case, a value closer to 0 results in better imputation [54], [55].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \quad (11)$$

In addition, DAC represents the technical ability to maintain the actual distribution of data values and was assessed in this study using the Kolmogorov-Smirnov Distance (D_{KS}). In cases where F_x and $F_{\hat{x}}$ are the empirical cumulative distribution function of x and $\hat{x}$, use equation (12) to calculate DKS, and a smaller distance value indicates a better interpolation result [55,56].

$$D_{KS} = \left\| F_x - F_{\hat{x}} \right\| \quad (12)$$

Subsequently, the complete dataset results' classification accuracy was analyzed, using several calibration algorithms, known as k-Nearest Neighbor (KNN). As shown in Figure 4, the KNN algorithm is selected based on

the review results of previous studies, where a classifier is usually used to evaluate the effectiveness of the interpolation algorithm.

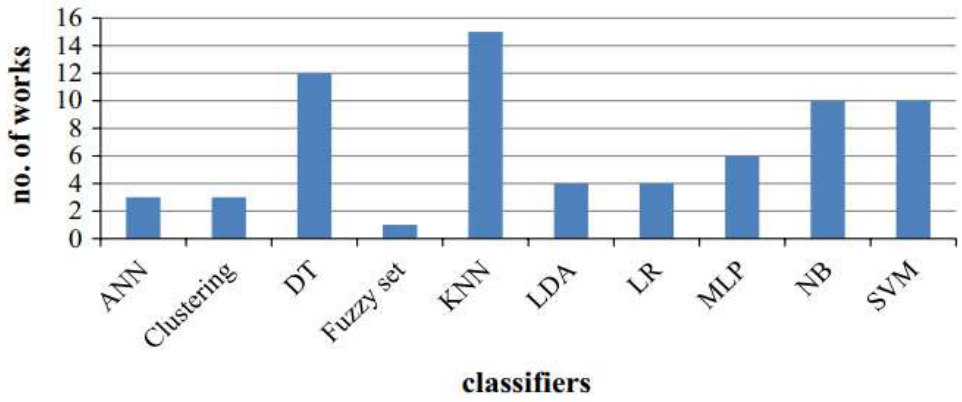


Fig 4. The Classifier used in previous studies [46].

3. Experimental Results

In this study, the first stage was selecting the Sonar dataset obtained from the UCI Machine Learning Repository (www.arsip.Ics.uci.edu/ml) and Kaggle Datasets (www.kaggle.com/datasets). Subsequently, a normality analysis was performed using the Shapiro-Wilk and Kolmogorov-Smirnov methods, to determine whether the data to be analyzed is normally distributed or not. In the analysis, data was regarded as normally distributed, in cases where a significance value above 0.05 (sig.> 0.05) is obtained. According to the test results, the sonar dataset data is not normally distributed. Figure 5 shows the next research stages.

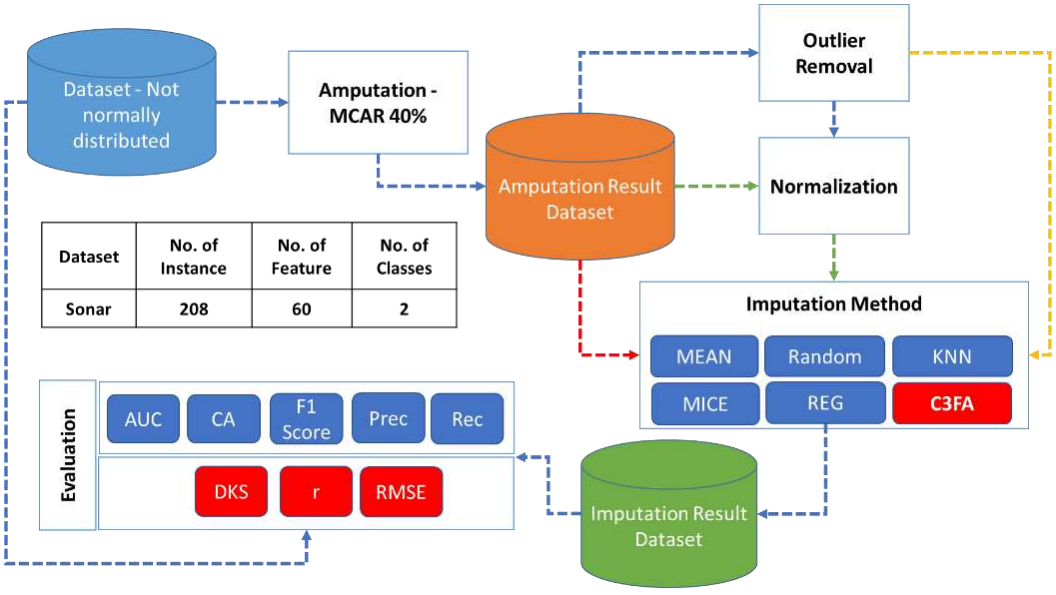


Fig 5. Proposed Method

Subsequently, the sonar dataset tested for normality is subjected to an amputation process, where 40% of the data is removed using the MCAR mechanism. A study by Schouten *et al*., (2018) show an important aspect of missing data research produces missing values in the complete data set, through the amputation procedure [58].

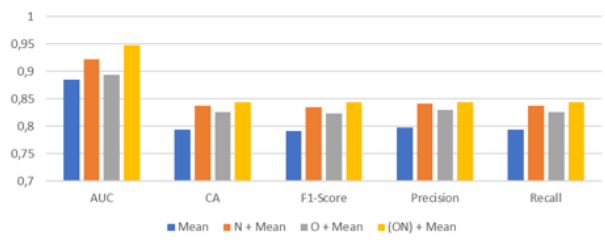
Algorithm 1. Generate missing values with R

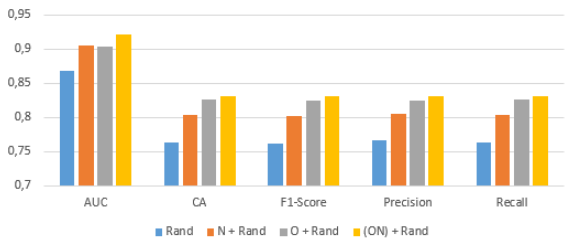
```
library(MASS)
library(VIM)
library(mice)
library(lattice)
library(readxl)
options(max.print = 1000)
ampute_sonar <- ampute(sonar, prop = 0.4,
                       patterns = c(0,1,1,1,1,1,0,1,1,1,1,0,1,1,1,1,
                                    0,1,1,1,1,0,1,1,1,1,0,1,1,1,1,0,
                                    1,1,1,1,0,1,1,1,1,0,1,1,1,1,0,1,
                                    1,1,1,0,1,1,1,1,0,1,1,1,1),
                       freq = NULL, mech = "MCAR", weights = NULL, std = TRUE,cont = TRUE,
                       type = NULL, odds = NULL, bycases = TRUE, run = TRUE)
```

The next stage is the imputation process, using 5 (five) standard methods, mean, random, multiple imputations, regression, and the proposed method conducted in the previous study, Class Center-Based Firefly Algorithm (C3-FA) [24]. To observe normalization and outlier detection's effect on the missing data imputation method, the simulation process was conducted in 4 ways:

1. Imputation;
2. Normalization + imputation;
3. Outlier removal's + imputation; and
4. Outlier removal's + normalization + imputation.

At this stage, several imputation methods, mean, random (Rand), linear regression (Reg), multiple (MI), KNN, and C3FA, are compared. Figure 6 shows the results of the evaluation using AUC, Accuracy, F1-Score, Precision, and Recall, with k-Nearest Neighbor (KNN), the most widely used classifier according to research.





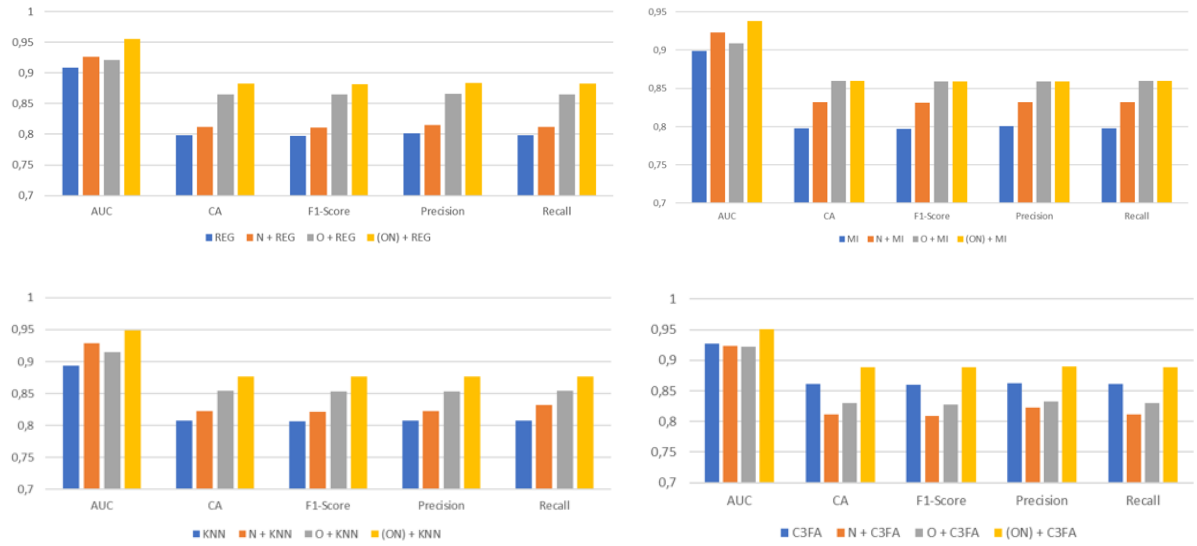


Fig 6. Evaluation results using AUC, Accuracy, F1-Score, Precision, and Recall.

Based on these experiments, outlier removal's (O) and normalization (N) before the imputation process had an effect on the evaluation results. Figure 7 shows a comparison of the evaluation results from six (6) imputation methods for the sonar dataset.

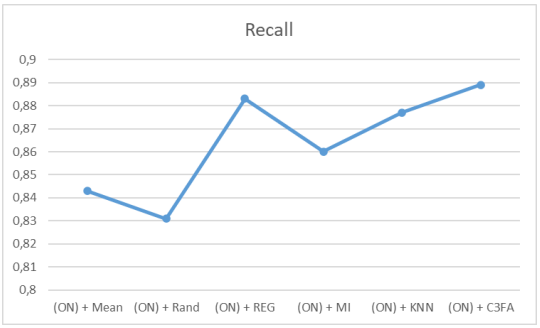


Fig 7. A Comparison of AUC, Accuracy, F1-Score, Precision, and Recall Results, from the ON + Imputation Method.

In addition to evaluations using AUC, Accuracy, F1-Score, Precision, and Recall, the proposed method is also evaluated based on the values of RMSE, DKS, and r. Figure 8 shows the result of this evaluation.

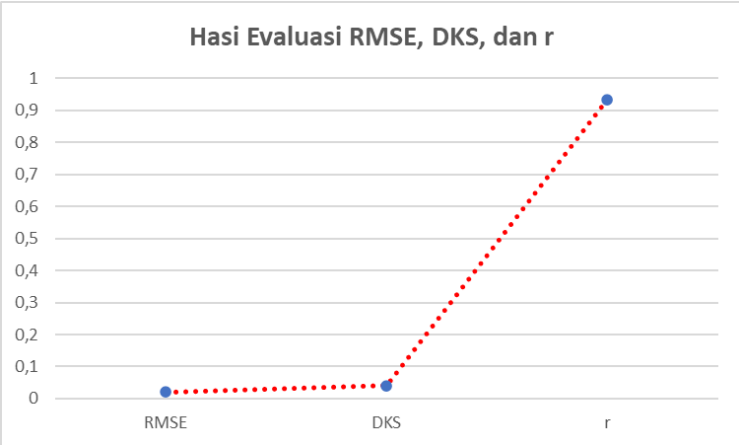


Fig 8. Evaluation of RMSE, DKS, and r (ON + C3FA) method

Based on the simulation results, the Pearson Correlation Coefficient (Pearson's r) value is close to 1, while the value Root Mean Squared Error (RMSE) is close to 0. This indicates the combination of outlier detection and normalization with the Firefly Algorithm in missing data handling based on class center (ON + C3FA), is an efficient technique for obtaining true data values.

5. Discussion and Conclusion

Adaptive search methods, such as those used in the Firefly algorithm, can be used to overcome missing values in a dataset. The initial objective feature of class center aids in the detection of the best imputation value. This corresponds to the fact that the Firefly algorithm is capable of determining the approximate value that is nearest to the known value. Combining normalization and imputation techniques has been shown in previous research to

improve accuracy values [32]. Meanwhile, other studies have emphasized the significance of detecting outliers in the observed dataset prior to imputation of missing values. [34].

Based on the simulation results using the sonar dataset and six (6) imputation methods, outlier removal (O) and normalization (N) before the imputation process were concluded to have an effect on the results. The simulation results with KNN classifier show accuracy, precision, F1-score, and recall are better, compared to the mean, a random value, linear regression, multiple imputation, KNN imputation, and C3-FA without outlier detection and normalization, prior to imputation, However, the comparison results show the proposed method, ON + C3-FA outperforms the others.

In order to obtain the true data value, integrating outlier detection and normalization in the Firefly Algorithm for handling missing data based on class center is an effective technique. This is indicated by the values of the Pearson's correlation coefficient (Pearson's r) and Root Mean Squared Error (RMSE) values being near to 1 and 0, respectively. In addition, the proposed method has the ability to maintain the true distribution of the data values as indicated by the average value of D_{KS} being close to 0.

ABBREVIATIONS

C3-FA: Class Center-Based Firefly Algorithm

RMSE: Root Mean Squared Error

EM: Expectation maximization

LR: Linear/logistic regression

LS: Least squares

DT: Decision Tree

KNN: k-Nearest Neighbor

RF: Random Forest

MAR: Missing at Random

MNAR: Missing Not at Random

MCAR: Missing Completely at Random

PAC: Predictive Accuracy

DAC: Distributional Accuracy

DECLARATIONS

Acknowledgments

We would like to thank Institut Teknologi Bandung and Telkom University for supporting this research.

Authors' contributions

The author confirms the sole responsibility for this manuscript fully as a sole author for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation. The author read and approved the final manuscript.

Funding

Not applicable. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials

The original dataset used for this study is available in:

1. UCI Machine Learning Repository (www.arsip.Ics.uci.edu/ml)
2. Kaggle Datasets (www.kaggle.com/datasets)

Competing interests

The author reports no potential conflict of interest.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Reference

1. Firmani D, Mecella M, Scannapieco M, Batini C. On the Meaningfulness of "Big Data Quality" (Invited Paper). Data Sci Eng. 2016;1:6–20.
2. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Medical Informatics and Decision Making [Internet]. 2016 [cited 2019 Apr 3];16. Available from: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-016-0318-z
3. Hayati Rezvan P, Lee KJ, Simpson JA. The rise of multiple imputation: a review of the reporting and implementation of the method in medical research. BMC Medical Research Methodology [Internet]. 2015 [cited 2019 Apr 9];15. Available from: http://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-015-0022-1
4. Kang H. The prevention and handling of the missing data. Korean Journal of Anesthesiology. 2013;64:402–6.
5. Ma Z, Chen G. Bayesian methods for dealing with missing data problems. Journal of the Korean Statistical Society. 2018;47:297–313.

6. Malarvizhi R, S. Thanamani A. K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation. IOSR Journal of Computer Engineering. 2012;6:12–5.

7. Marlin BM. Missing Data Problems in Machine Learning [Internet]. Department of Computer Science, University of Toronto; 2008. Available from: https://people.cs.umass.edu/~marlin/research/phd_thesis/marlin-phd-thesis.pdf

8. Ng CG, Yusoff MSB. Missing Values in Data Analysis: Ignore or Impute? Education in Medicine Journal [Internet]. 2011 [cited 2019 Apr 8];3. Available from: http://eduimed.usm.my/EIMJ20110301/EIMJ20110301_02.pdf

9. Pampaka M, Hutcheson G, Williams J. Handling missing data: analysis of a challenging data set using multiple imputation. International Journal of Research & Method in Education. 2016;39:19–37.

10. Rahman MdG, Islam MZ. Missing value imputation using a fuzzy clustering-based EM approach. Knowledge and Information Systems. 2016;46:389–422.

11. Zhao Y, Long Q. Multiple imputation in the presence of high-dimensional data. Statistical Methods in Medical Research. 2016;25:2021–35.

12. Armina R, Mohd Zain A, Ali NA, Sallehuddin R. A Review on Missing Value Estimation Using Imputation Algorithm. Journal of Physics: Conference Series. 2017;892:012004.

13. Cao L. Data science thinking. New York, NY: Springer Science+Business Media; 2018.

14. Nishanth KJ, Ravi V. Probabilistic neural network based categorical data imputation. Neurocomputing. 2016;218:17–25.

15. Van Hulse J, Khoshgoftaar TM. Incomplete-case nearest neighbor imputation in software measurement data. Information Sciences. 2014;259:596–610.

16. Nugroho H, Surendro K. Missing Data Problem in Predictive Analytics. 8th International Conference on Software and Computer Applications - ICSCA '19 [Internet]. Penang, Malaysia: ACM Press; 2019 [cited 2019 Aug 3]. p. 95–100. Available from: http://dl.acm.org/citation.cfm?doid=3316615.3316730

17. Jugulum R. Importance of Data Quality for Analytics. In: Sampaio P, Saraiva P, editors. Quality in the 21st Century [Internet]. Cham: Springer International Publishing; 2016 [cited 2019 Apr 8]. p. 23–31. Available from: http://link.springer.com/10.1007/978-3-319-21332-3_2

18. Wazurkar P, Bhadoria RS, Bajpai D. Predictive analytics in data science for business intelligence solutions. 2017 7th International Conference on Communication Systems and Network Technologies (CSNT) [Internet]. Nagpur: IEEE; 2017 [cited 2019 Apr 8]. p. 367–70. Available from: https://ieeexplore.ieee.org/document/8418568/

19. Deb R, Liew AW-C. Missing value imputation for the analysis of incomplete traffic accident data. Information Sciences. 2016;339:274–89.

20. Farhangfar A, Kurgan L, Dy J. Impact of imputation of missing values on classification error for discrete data. Pattern Recognition. 2008;41:3692–705.

21. Pedersen A, Mikkelsen E, Cronin-Fenton D, Kristensen N, Pham TM, Pedersen L, et al. Missing data and multiple imputation in clinical epidemiological research. Clinical Epidemiology. 2017;Volume 9:157–66.

22. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR, Verleysen M. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing. 2009;72:1483–93.

23. Dong Y, Peng C-YJ. Principled missing data methods for researchers. SpringerPlus. 2013;2:222.

24. Bhati S, Kumar Gupta MKG. Missing Data Imputation for Medical Database: Review. International Journal of Advanced Research in Computer Science and Software Engineering. 2016;6.

25. Nugroho H, Utama NP, Surendro K. Class center-based firefly algorithm for handling missing data. J Big Data. 2021;8:37.

26. Nugroho H, Utama NP, Surendro K. Performance Evaluation for Class Center-Based Missing Data Imputation Algorithm. Proceedings of the 2020 9th International Conference on Software and Computer Applications [Internet]. Langkawi Malaysia: ACM; 2020 [cited 2021 Jan 15]. p. 36–40. Available from: https://dl.acm.org/doi/10.1145/3384544.3384575

27. Alizadeh Naeini A, Babadi M, Homayouni S. Assessment Of Normalization Techniques On The Accuracy Of Hyperspectral Data Clustering. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2017;XLII-4/W4:27–30.

28. Huang H-C, Qin L-X. Empirical evaluation of data normalization methods for molecular classification. PeerJ. 2018;6:e4584.

29. KumarSingh B, Verma K, S. Thoke A. Investigations on Impact of Feature Normalization Techniques on Classifier& Performance in Breast Tumor Classification. IJCA. 2015;116:11–5.

30. Rozenstein O, Paz-Kagan T, Salbach C, Karnieli A. Comparing the Effect of Preprocessing Transformations on Methods of Land-Use Classification Derived From Spectral Soil Measurements. IEEE J Sel Top Appl Earth Observations Remote Sensing. 2015;8:2393–404.

31. Alshdaifat E, Alshdaifat D, Alsarhan A, Hussein F, El-Salhi SMFS. The Effect of Preprocessing Techniques, Applied to Numeric Features, on Classification Algorithms' Performance. Data. 2021;6:11.

32. Madhu G, Lalith Bharadwaj B, Sai Vardhan K, Naga Chandrika G. A Normalized Mean Algorithm for Imputation of Missing Data Values in Medical Databases. In: Saini HS, Singh RK, Tariq Beg M, Sahambi JS, editors. Innovations in Electronics and Communication Engineering [Internet]. Singapore: Springer Singapore; 2020 [cited 2021 Apr 1]. p. 773–81. Available from: http://link.springer.com/10.1007/978-981-15-3172-9_72

33. Christobel A, Prakasam S. The Negative Impact of Missing Value Imputation in Classification of Diabetes Dataset and Solution for Improvement. IOSRJCE. 2012;7:16–23.

34. Huang M-W, Lin W-C, Tsai C-F. Outlier Removal in Model-Based Missing Value Imputation for Medical Datasets. Journal of Healthcare Engineering. 2018;2018:1–9.

35. Garcia S, Derrac J, Cano JR, Herrera F. Prototype Selection for Nearest Neighbor Classification: Taxonomy and Empirical Study. IEEE Trans Pattern Anal Mach Intell. 2012;34:417–35.

36. Leyva E, González A, Pérez R. Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective. Pattern Recognition. 2015;48:1523–37.

37. Wada K. Outliers in official statistics. Jpn J Stat Data Sci. 2020;3:669–91.

38. Kim M-G, Shin K-I. A Multiple Imputation for Reducing Outlier Effect. The Korean Journal of Applied Statistics. 2014;27:1229–41.

39. Branden KV, Verboven S. Robust data imputation. Computational Biology and Chemistry. 2009;33:7–13.

40. Toka O, ÇetiN M. Imputation and Deletion Methods Under The Presence of Missing Values and Outliers: A Comparative Study. Gazi University Journal of Science [Internet]. 2016;29. Available from: https://dergipark.org.tr/tr/download/article-file/273154

41. Cheng T-C, Victoria-Feser M-P. High-breakdown estimation of multivariate mean and covariance with missing observations. British Journal of Mathematical and Statistical Psychology. 2002;55:317–35.

42. Hubert M, Rousseeuw PJ, Vanden Branden K. ROBPCA: A New Approach to Robust Principal Component Analysis. Technometrics. 2005;47:64–79.

43. Kumar N, Hoque MdA, Shahjaman Md, Islam SMS, Mollah MdNH. A New Approach of Outlier-robust Missing Value Imputation for Metabolomics Data Analysis. CBIO. 2018;14:43–52.

44. García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR. Pattern classification with missing data: a review. Neural Computing and Applications. 2010;19:263–82.

45. Peng L, Lei L. A Review of Missing Data Treatment Methods. Int Journal of Intel Inf Manag Syst Tech. 2005;8.

46. Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). Artif Intell Rev. 2020;53:1487–509.

47. Hu L-Y, Huang M-W, Ke S-W, Tsai C-F. The distance function effect on k-nearest neighbor classification for medical datasets. SpringerPlus. 2016;5:1304.

48. Nugroho H, Utama NP, Surendro K. Comparison Method for Handling Missing Data in Clinical Studies. 9th International Conference on Software and Computer Applications (ICSCA). Langkawi, Malaysia; 2020. p. 6.

49. Pan R, Yang T, Cao J, Lu K, Zhang Z. Missing data imputation by K nearest neighbours based on grey relational structure and mutual information. Applied Intelligence. 2015;43:614–32.

50. Pires IM, Hussain F, Garcia NM, Zdravevski E. Improving Human Activity Monitoring by Imputation of Missing Sensory Data: Experimental Study. Future Internet. 2020;12:155.

51. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol. 2017;70:407.

52. Quintano C, Castellano R, Rocca A. Influence of Outliers on Some Multiple Imputation Methods. Metodološki zvezki. 2010;7:16.

53. Kulkarni A, Chong D, Batarseh FA. Foundations of data imbalance and solutions for a data democracy. Data Democracy [Internet]. Elsevier; 2020 [cited 2021 Mar 31]. p. 83–106. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780128183663000058

54. Yuliansyah H, Othman ZA, Bakar AA. Taxonomy of Link Prediction for Social Network Analysis: A Review. IEEE Access. 2020;8:183470–87.

55. Pompeu Soares J, Seoane Santos M, Henriques Abreu P, Araújo H, Santos J. Exploring the Effects of Data Distribution in Missing Data Imputation. In: Duivesteijn W, Siebes A, Ukkonen A, editors. Advances in Intelligent Data Analysis XVII [Internet]. Cham: Springer International Publishing; 2018 [cited 2019 May 29]. p. 251–63. Available from: http://link.springer.com/10.1007/978-3-030-01768-2_21

56. Santos MS, Soares JP, Henriques Abreu P, Araújo H, Santos J. Influence of Data Distribution in Missing Data Imputation. In: ten Teije A, Popow C, Holmes JH, Sacchi L, editors. Artificial Intelligence in Medicine [Internet]. Cham: Springer International Publishing; 2017 [cited 2019 May 29]. p. 285–94. Available from: http://link.springer.com/10.1007/978-3-319-59758-4_33

57. Oytun M, Tinazci C, Sekeroglu B, Acikada C, Yavuz HU. Performance Prediction and Evaluation in Female Handball Players Using Machine Learning Models. IEEE Access. 2020;8:116321–35.

58. Schouten R. Generating missing values for simulation purposes: a multivariate amputation procedure. Journal of Statistical Computation and Simulation. 2018;88:2909–30.

Figures

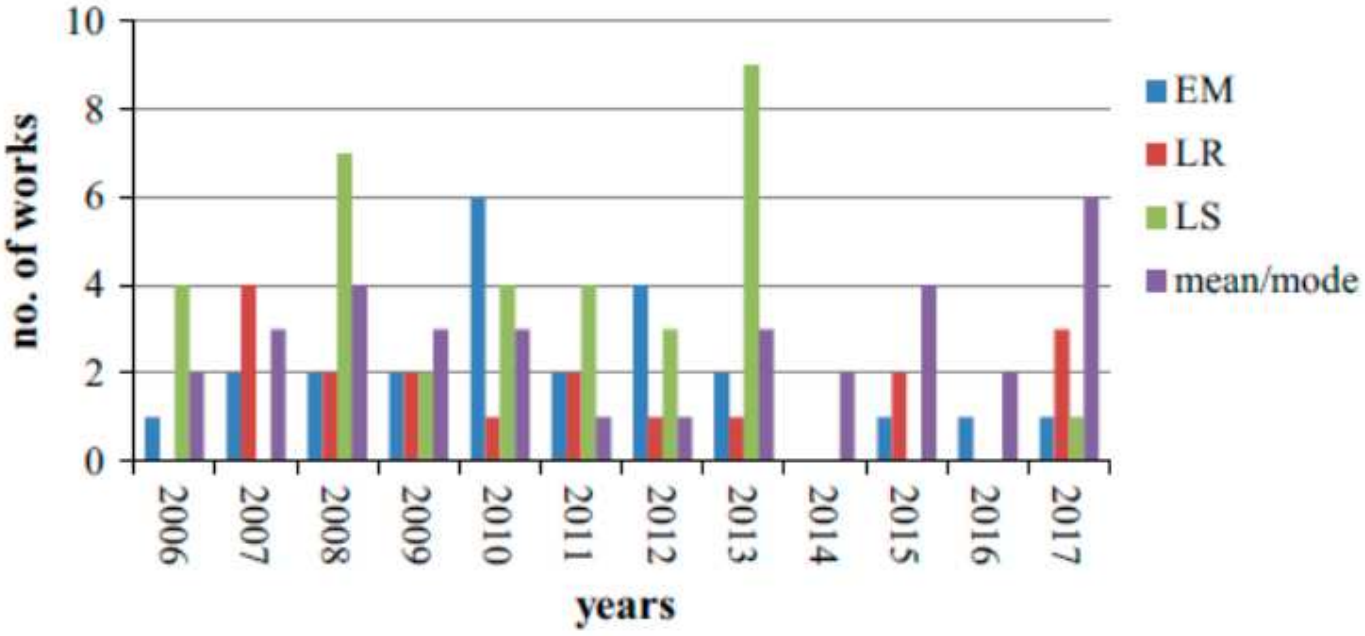


Figure 1

Distribution of the number of studies using EM, LR, LS, and mean/mode techniques [46]

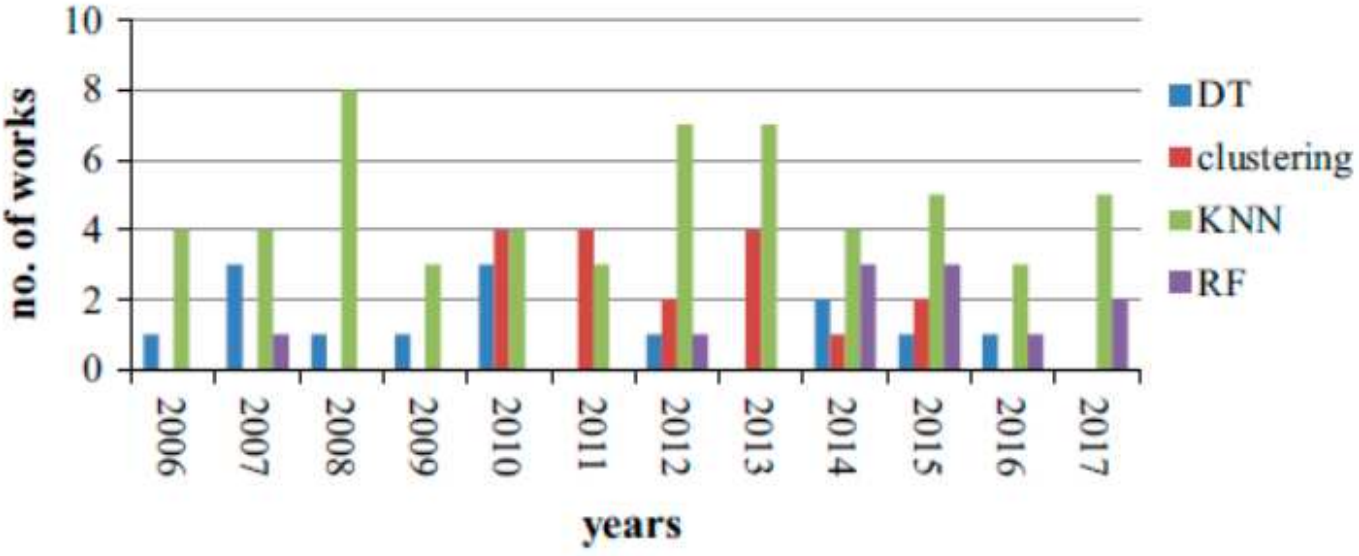


Figure 2

Distribution of the number of studies using DT, clustering, KNN, and RF techniques [46]

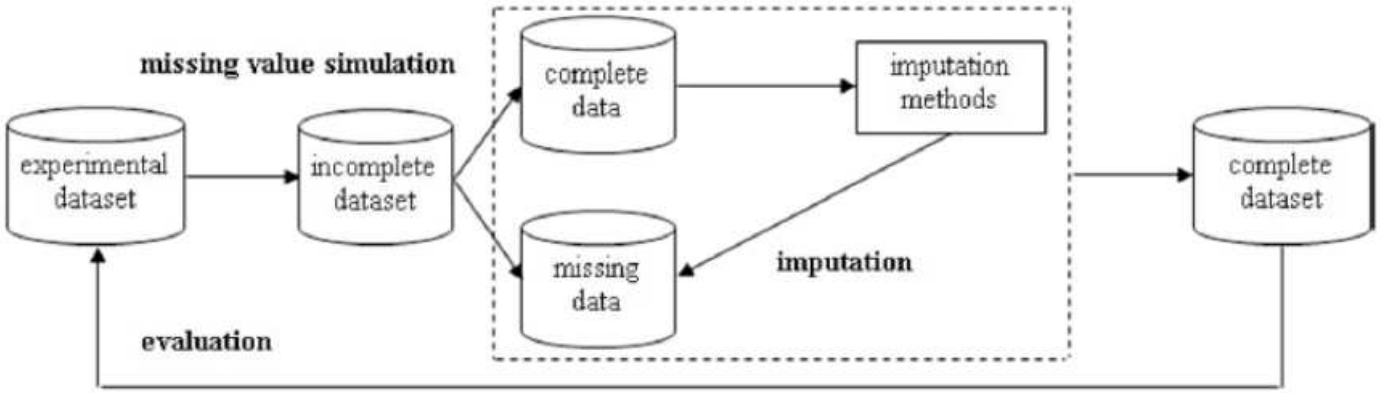


Figure 3

The experimental design procedure for missing data imputation experimental [46].

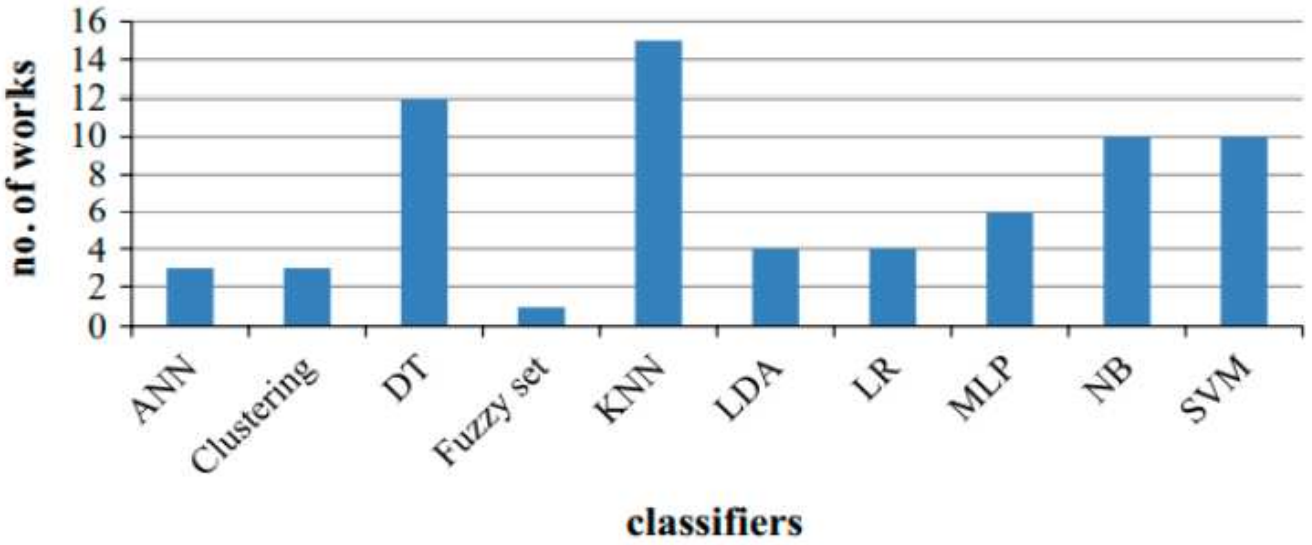


Figure 4

The Classifier used in previous studies [46].

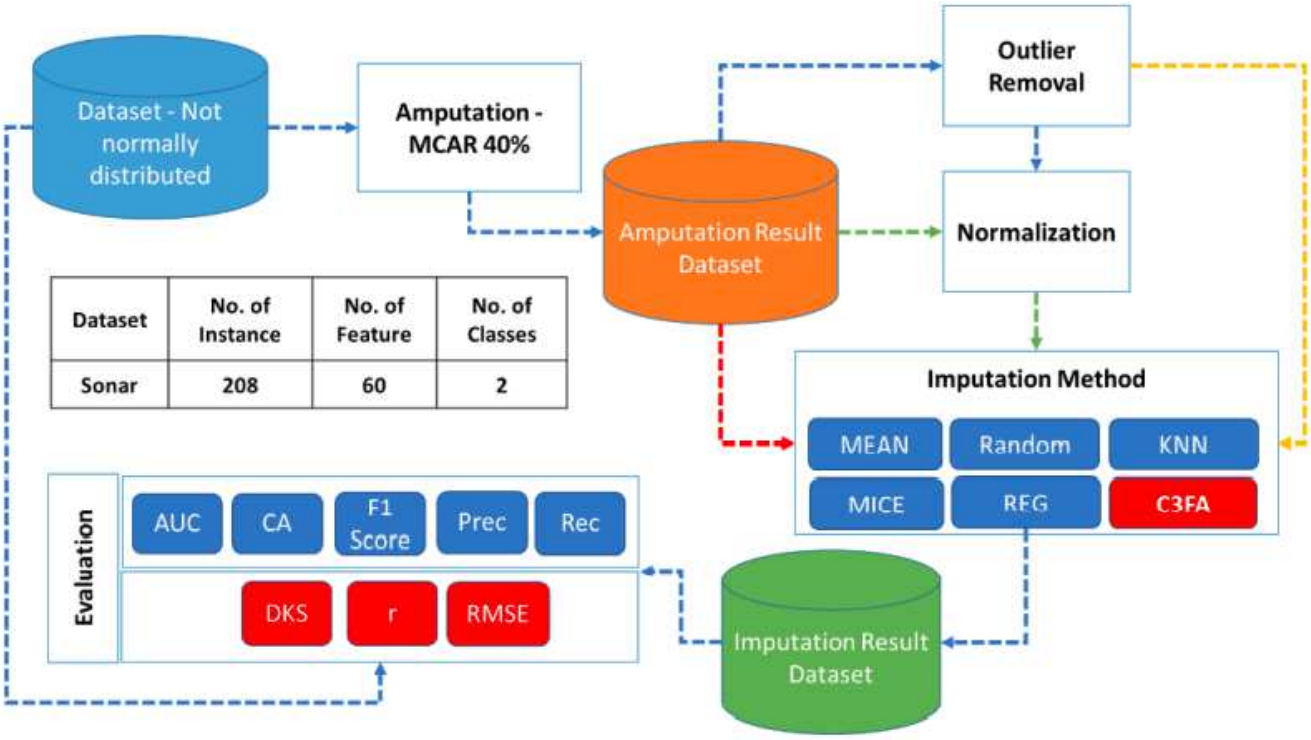


Figure 5

Proposed Method

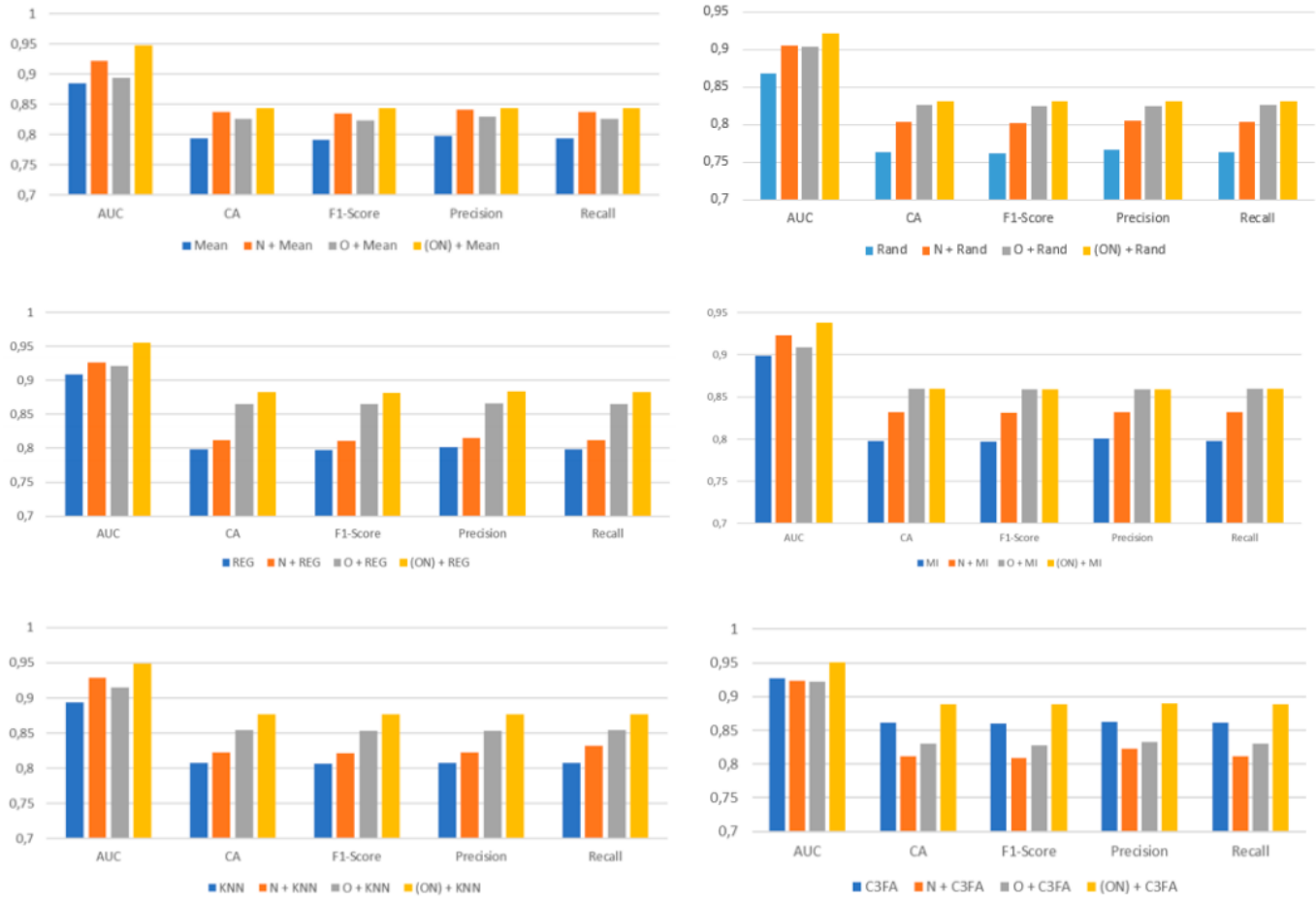


Figure 6

Evaluation results using AUC, Accuracy, F1-Score, Precision, and Recall.

Figure 7

A Comparison of AUC, Accuracy, F1-Score, Precision, and Recall Results, from the ON + Imputation Method.

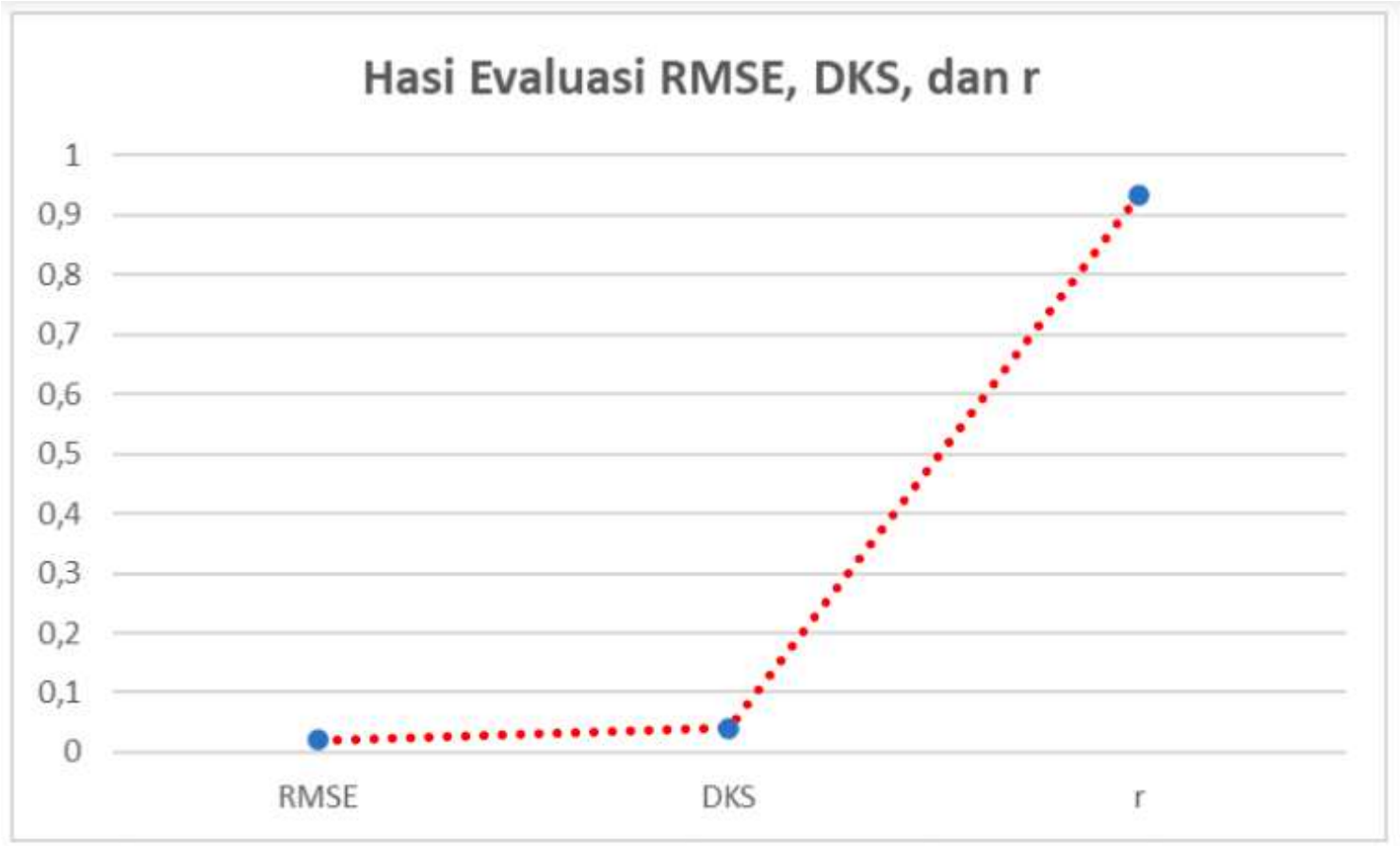


Figure 8

Evaluation of RMSE, DKS, and r (ON + C3FA) method