

## THE EFFECT OF REGULARIZED REGRESSION AND TREE-BASED MISSING DATA IMPUTATION METHODS ON CLASSIFICATION PERFORMANCE IN HIGH DIMENSIONAL DATA

Buğra VAROL<sup>1\*</sup>, İmran KURT ÖMÜRLÜ<sup>2</sup>, Mevlüt TÜRE<sup>2</sup>

<sup>1</sup>Adnan Menderes University, Institute of Health Sciences, Division of Biostatistics, 09010, Aydın, Türkiye


<sup>2</sup>Adnan Menderes University, Faculty of Medicine, Division of Biostatistics, 09010, Aydın, Türkiye


**Abstract:** Missing data is an important problem in the analysis and classification of high dimensional data. The aim of this study is to compare the effects of four different missing data imputation methods on classification performance in high dimensional data. In this study, missing data imputation methods were evaluated using data sets, whose independent variables between mixed correlated with each other, for binary dependent variable,  $p=500$  independent variables,  $n=150$  units and 1000 times running simulation. Missing data structures were created according to different missing rates. Different datasets were obtained by imputing the missing values using different methods. Regularized regression methods such as least absolute shrinkage and selection operator (lasso) and elastic net regression were used for imputation, as well as tree-based methods such as support vector machine and classification and regression trees. At the end of simulation, the classification scores of the methods were obtained by gradient boosting machine and the missing data prediction performances were evaluated according to the distance of these scores from the reference. Our simulation demonstrates that regularized regression methods outperform tree-based methods in classifying high dimensional datasets. Additionally, it was found that the increase in the amount of missing values reduced the classification performance of the methods in high dimensional data.


**Keywords:** Gradient boosting machine, High dimensional data, Imputation, Classification, Simulation

\*Corresponding author: Adnan Menderes University, Institute of Health Sciences, Division of Biostatistics, 09010, Aydın, Türkiye

E mail: bugravarol87@gmail.com (B. VAROL)

Buğra VAROL  <https://orcid.org/0000-0001-8052-7782>

İmran KURT ÖMÜRLÜ  <https://orcid.org/0000-0003-2887-6656>

Mevlüt TÜRE  <https://orcid.org/0000-0003-3187-2322>

Received: August 12, 2024

Accepted: October 21, 2024

Published: November 15, 2024

**Cite as:** Varol B, Kurt Ömürlü İ, Türe M. 2024. The effect of regularized regression and tree-based missing data imputation methods on classification performance in high dimensional data. *BSJ Eng Sci*, 7(6): 1263-1269.

### 1. Introduction

Dealing with missing data is a crucial aspect of statistical analysis. In statistical studies, missing values occur when observations for a variable cannot be obtained due to various reasons (Jadhav et al., 2019). The presence of missing data is a common issue in clinical research and can significantly affect the data analysis process. Understanding the source and structure of missing data is crucial. Naive analyses, such as complete-case and available-case analysis, can cause bias, loss of efficiency, and unreliable results (Enders, 2022). When dealing with missing data, it is advisable to make use of any available partial information to estimate the missing values and analyze the complete dataset. This approach is more preferable than excluding the missing units from the analysis, as it helps to preserve the integrity and completeness of the data (Rubin, 1988; Little and Rubin, 2019).

Missing data, which negatively affects statistical analysis processes, is also an important problem for researchers dealing with classification problems in high dimensional data. When training a model, several commonly used classification methods are unable to deal with missing values in the training data (Deng et al., 2016). Therefore, it has become imperative to research and develop appropriate imputation methods for missing values in

the training data to enhance the overall performance of the classifier on test data. Many of the methods that handle missing data are not suitable for high dimensional data because of their theoretical structure. Tree-based and regularized regression-based methods are remarkable in studies on missing values in high dimensional data (Yin et al., 2016; Zhao and Long, 2016). The objective of this research is to assess the impact of regularized regression imputation methods, such as least absolute shrinkage and selection operator (lasso) and elastic net regression, and tree-based missing data imputation methods, such as support vector machine (SVM) and classification and regression trees (CART) on the classification performance by gradient boosting machine (GBM) in simulated high dimensional data.

### 2. Materials and Methods

In this section, we first described the theory of imputation methods and GBM structure. At the end of this section, we gave information about the simulation algorithm.

#### 2.1. Regularized Regression Models

Regularized regression is a statistical technique that is quite similar to ordinary regression, whether it is linear or logistic. The main difference between the two lies in the fact that regularized regression adds an extra



constraint, which has the objective of shrinking the values of unimportant regression coefficients towards zero. This technique is particularly useful when dealing with high dimensional datasets. In such situations, regularized regression helps in avoiding overfitting, which can occur when the model is too complex and captures the noise present in the data. By shrinking the coefficients, regularized regression allows the model to focus on the most relevant predictors, thereby improving its generalization performance on new, unseen data (Przednowek and Wiktorowicz, 2013; Patil and Kim, 2020).

**2.1.1. Lasso regression**

Lasso regression is a technique used in linear regression to reduce the complexity of models when working with high dimensional datasets. The lasso regression, also known as L1 regularization, works by adding a penalty term to the cost function. This penalty shrinks some variable coefficients to zero, selecting only the most significant features, and induces a sparse solution (Breiman, 1995; Tibshirani, 1996).

**2.1.2. Elastic net regression**

Elastic net regression is a general regularization technique that combines L1 and L2 regularization techniques for feature selection and reduction. The regularization term L2, also known as Ridge regression, suggests keeping the coefficients small but not zero, which in turn reduces the coefficients of less significant features. Thus, some coefficients shrink to exactly zero, while others shrink towards each other. This allows for variable selection and reduces the complexity of the model (Zou and Hastie, 2005; Friedman et al., 2010).

**2.2. Tree-Based Models**

Tree-based models are a type of machine learning algorithm that can also be applied for high dimensional data and fall under the category of nonparametric models. These models operate by dividing the feature space into smaller, non-overlapping regions. The partitioning process is done in such a way that the response values within each region are similar to each other. Overall, tree-based models are a powerful tool for solving a wide range of supervised learning problems, including regression and classification tasks (Chang and Chen, 2005; Clark and Pregibon, 2017).

**2.2.1. Support vector machine**

SVM is a popular machine learning algorithm used for classification and regression analysis. SVM works by finding the best possible boundary that separates data points into different classes. It tries to maximize the margin between the classes, which is the distance between the boundary and the closest data points. SVM is particularly useful when working with high dimensional datasets and can handle both linear and non-linear data. It is also known for its ability to deal with noisy data and outliers (Cortes and Vapnik, 1995; Hastie et al., 2009).

**2.2.2. Classification and regression trees**

CART is a decision tree algorithm that can be utilized for both classification and regression tasks. The data is

divided into subsets by this algorithm in a recursive manner, depending on feature values. This process continues until a stopping criterion, such as reaching a maximum depth or minimum number of samples per leaf node, is met. At each split, the algorithm determines the feature that can separate the data into various classes or produce the smallest residual sum of squares for regression (Breiman, 2017; Loh, 2011).

**2.3. Imputation Algorithm for Dealing with Missing Data**

Let's consider a data set  $X$  consisting of  $n$  rows and  $p$  columns, where each row represents an observation and each column represents a variable denoted by  $x_1, \dots, x_p$ . We assume that the first  $t$  variables have missing. We use the notation  $x_{obs,j}$  to represent the observed components and  $x_{mis,j}$  to represent the missing components for variable  $j$  where  $n_{obs,j}$  and  $n_{mis,j}$  represent the number of samples, respectively. The collection of  $p - 1$  variables in  $X$ , excluding  $x_j$ , can be denoted as  $X_{-j} = (x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_t, x_{t+1}, \dots, x_p)$ . Let  $X_{obs,-j}$  and  $X_{mis,-j}$  be the components of  $X_{-j}$  that correspond to the complement data of  $x_{obs,j}$  and  $x_{mis,j}$  (Deng et al., 2016; Stekhoven and Bühlmann, 2012).

All the imputation methods employed in the study run based on a particular algorithm. The algorithm can be described as follows (Zhang, 2016; Zhang et al., 2021):

1. Sort  $X$  in descending order based on the amount of missing values.
2. Use mean imputation to make an initial guess for any missing values and update  $X$  matrix ( $\hat{X}$ ).
3. Fit a model ( $x_{obs,j} \sim \hat{X}_{obs,-j}$ ).
4. Predict  $x_{mis,j}$  using  $\hat{X}_{mis,-j}$  and obtain  $\hat{x}_{mis,j}$ .
5. Update variable  $j$  of  $\hat{X}$  ( $\hat{X} \leftarrow \hat{x}_j$ ).
6. Repeat steps 3-5 for  $j = 1, 2, \dots, t$ .
7. Obtain ultimate imputed dataset ( $\hat{X}^{new} \leftarrow \hat{X}$ ).

**2.4. Gradient Boosting Machine**

GBM is a type of machine learning algorithm that uses an ensemble of decision trees to make predictions. It is a powerful and popular method for both regression and classification tasks, and is known for its ability to handle complex datasets with high accuracy (Schapire, 2003; Tian et al., 2020). GBM algorithm works by iteratively adding decision trees to a model, with each subsequent tree focusing on the errors made by the previous tree. This allows the model to gradually improve its performance over time, leading to highly accurate predictions (Nawar and Mouazen, 2017; Zhang et al., 2019). Any arbitrary loss function  $L(\cdot, \cdot)$  can be used here.

Let  $n$  and  $p$  indicate the number of observations and independent variables, respectively;  $y_i \in R^a$  denotes the dependent variable value of each observation ( $a=1$  for regression,  $a=2$  for classification) and  $\{(x_i, y_i) | x_i \in R^p, y_i \in R^a\}_{i=1}^n$  denotes the training set. According to (Elith et al., 2008) the GBM algorithm can be described as follows:

1. To begin, the weak classifiers must be initialized by solving the equation 1 below, where  $\gamma$  denotes the step size:

$$F_0(x) = \operatorname{argmin}_{\gamma} \sum_{i=1}^n L(y_i, \gamma) \quad (1)$$

2. Starting from first iteration  $m = 1$ , and up to a maximum of  $M$  iterations for learning:

a. The pseudo-residuals is computed for  $i = 1, 2, \dots, n$  as follows (equation 2):

$$r_{im} = - \left[ \frac{\partial L[y_i, F_{m-1}(x_i)]}{\partial F_{m-1}(x_i)} \right] \quad (2)$$

b. We need to train a new base model  $h_m(x_i)$  using the revised dataset  $\{x_i, r_{im}\}_{i=1}^n$ . Then the parameter  $\gamma_m$  is defined to solve the optimization problem as follow (equation 3):

$$\gamma_m = \operatorname{argmin}_{\gamma} \sum_{i=1}^n L[y_i, F_{m-1}(x_i) + \gamma h_m(x_i)] \quad (3)$$

c. At last, we get our final strong classifier (equation 4):

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (4)$$

### 2.5. Performance Evaluation Criteria

The performance of missing data imputation methods in predicting original values and their impact on classification was assessed using the area under the receiver operating characteristic (ROC) curve (AUC) and F1 score. AUC measures model separability, ranging from 0 to 1, with higher values indicating better performance. (Hanley and McNeil, 1982; Fawcett, 2006). The F1 score, which balances precision and recall, is particularly useful when false positives and false negatives carry similar consequences, with higher scores indicating a more balanced model (Tharwat, 2021).

### 2.6. Simulation

Statistical analysis of the study was performed using R software, version 4.2.3 (R Foundation for Statistical Computing, Vienna, Austria). Most existing MI methods rely on the assumption of missing at random (MAR), i.e., missingness only depends on observed data; our work also focuses on MAR. We set the sample size to  $n=150$  and included  $p=500$  predictors in simulated dataset. The dataset contains a binary outcome  $y$ , which is fully observed, and  $X = (X_1, X_2) = (x_1, \dots, x_p)$ . Firstly, a set of  $p - 10$  independent variables denoted as  $X_1 = (x_{11}, \dots, x_p)$  were created by drawing from a multivariate standard normal distribution with a mean vector of  $(0, \dots, 0)_{p-10}$  and the correlation matrix was defined as the absolute values of its all off-diagonal terms were a maximum of 0.7. We randomly selected 50 variables ( $X_s$ ) from  $X_1$ . We also generated a separate group of 10 variables called  $X_2 = (x_1, \dots, x_{10})$  from a normal distribution ( $X_2 \sim N_{p_2}(\mu_2, \sigma_2^2)$ ), where  $\mu_2$  is a linear combination of  $X_s$  and  $\sigma_2^2$  is equal to 1.5. To produce a

binary outcome, certain values were randomly generated and divided into two groups depending on whether they were below or above the median value. These values were generated from a normal distribution ( $N_p(\mu, \sigma^2)$ ), where  $\mu$  denotes linear combination of  $X$  and  $\sigma^2$  is equal to 5. Then missing values with MAR mechanism were created in  $(x_1, \dots, x_{10})$ , resulting in approximately 10%, 20%, 30%, 40% and 50% missing rates per variable. Thus, datasets with different missing rates were obtained. Missing values were imputed using lasso, elastic net, SVM and CART. To assess the classification performance of the methods, complete (reference) and imputed datasets were initially split into training and test subsets using a 70:30 ratio, selected randomly. The models were trained using training sets, while test sets were utilized to obtain AUC and F1 values of the models. Missing data prediction performances were evaluated according to the distance of these values from the reference. The processes were repeated 1000 times..

## 3. Results

The performance evaluation of the imputation models was done by AUC and F1 values. Performance metrics were calculated as median (25th – 75th percentiles) and presented visually using forest plots. Hence, the evaluation process identified the methods that exhibited comparable performance and generated results in proximity to the ones derived from the reference dataset. The reference data set provided median values of 0.945 and 0.902 for AUC and F1 values, respectively. The imputed datasets had the following median ranges for AUC and F1: 0.945-0.946 and 0.900-902 for the 10% missing rate; 0.943-0.944 and 0.898-900 for the 20% missing rate; 0.931-0.940 and 0.884-897 for the 30% missing rate; 0.928-0.937 and 0.885-889 for the 40% missing rate; 0.903-0.927 and 0.857-880 for the 50% missing rate. The performance of imputation methods was also assessed based on their AUC and F1 values. The median ranges for these values were as follows respectively: for the lasso, 0.925-0.945 and 0.880-0.900; for the elastic net, 0.927-0.946 and 0.880-0.900; for the SVM, 0.903-0.945 and 0.857-0.902; for the CART, 0.908-0.945 and 0.868-0.898 (Table 1).

After evaluating the effectiveness of all imputation techniques, considering AUC and F1 values, it was noticed that, all methods demonstrated good performance and were very close to the reference at 10% and 20% missing rates. However, the performance of the methods began to decline after the 20% missing rate. The regularized regression models performed slightly better than tree-based methods at the missing rate of 30%. Although the performance of the SVM method did not change at the missing rate of 40%, the performance of other methods decreased. However, regularized regression methods maintained their superiority over tree-based methods. As the rate of missing data increased to 50%, the differences in performance between

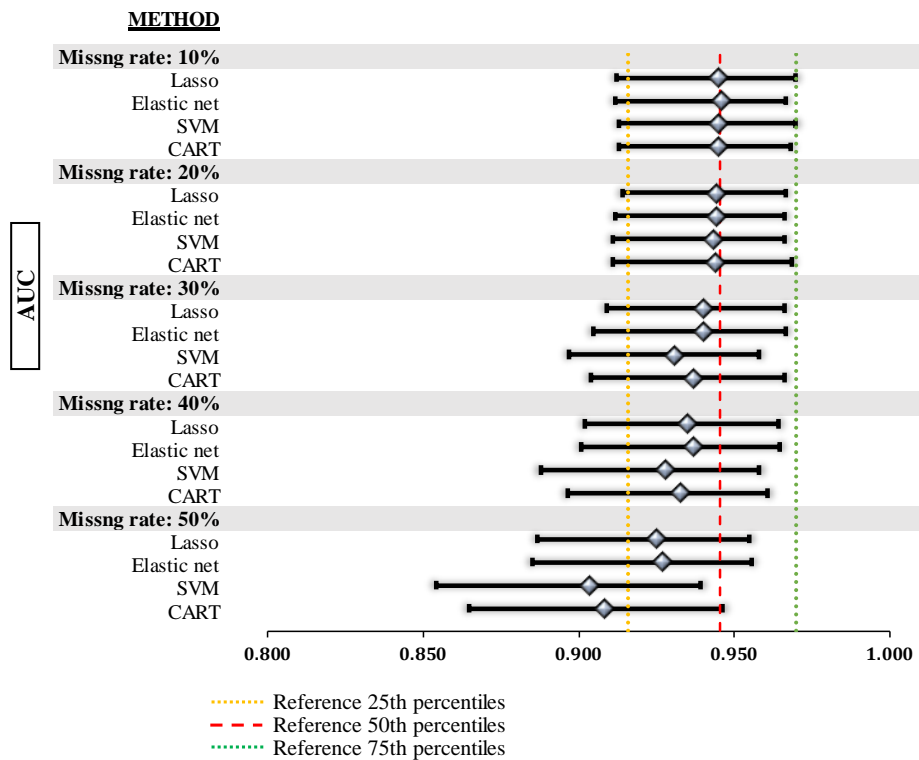
regularized regression methods and tree-based methods became more pronounced, with the former being closer to the reference than the latter. As a result, the tree-based methods moved further away from the reference as the rate of missing data increased, while the regularized regression methods showed more robustness in handling the missing data and produced better results (Figures 1 and 2).

values for all missing rates, was applied to determine the relationships among the methods and which methods were close to the reference. As shown the dendrogram graph in Figure 3, it was observed that the lasso and elastic net methods clustered together with the reference. However, the SVM and CART methods have formed a distinct cluster, separate from the others (Figure 3).

Hierarchical clustering analysis, based on AUC and F1

**Table 1.** Classification performances of the imputation models according to varying missing rates

		Missing Rate				
Method		%10	%20	%30	%40	%50
AUC	Reference	0.945 (0.916 - 0.970)	0.945 (0.916 - 0.970)	0.945 (0.916 - 0.970)	0.945 (0.916 - 0.970)	0.945 (0.916 - 0.970)
	Lasso	0.945 (0.912 - 0.970)	0.944 (0.914 - 0.966)	0.940 (0.909 - 0.966)	0.935 (0.902 - 0.964)	0.925 (0.887 - 0.955)
	Elastic net	0.946 (0.912 - 0.966)	0.944 (0.912 - 0.966)	0.940 (0.905 - 0.966)	0.937 (0.901 - 0.964)	0.927 (0.885 - 0.956)
	SVM	0.945 (0.913 - 0.970)	0.943 (0.911 - 0.966)	0.931 (0.897 - 0.958)	0.928 (0.888 - 0.958)	0.903 (0.854 - 0.939)
	CART	0.945 (0.913 - 0.968)	0.944 (0.911 - 0.968)	0.937 (0.904 - 0.966)	0.933 (0.897 - 0.960)	0.908 (0.865 - 0.946)
F1	Reference	0.902 (0.865 - 0.933)	0.902 (0.865 - 0.933)	0.902 (0.865 - 0.933)	0.902 (0.865 - 0.933)	0.902 (0.865 - 0.933)
	Lasso	0.900 (0.864 - 0.933)	0.900 (0.865 - 0.932)	0.897 (0.857 - 0.929)	0.889 (0.850 - 0.923)	0.880 (0.837 - 0.915)
	Elastic net	0.900 (0.864 - 0.930)	0.900 (0.864 - 0.931)	0.895 (0.857 - 0.929)	0.889 (0.852 - 0.929)	0.880 (0.837 - 0.913)
	SVM	0.902 (0.864 - 0.933)	0.898 (0.857 - 0.930)	0.884 (0.846 - 0.917)	0.885 (0.842 - 0.919)	0.857 (0.815 - 0.897)
	CART	0.898 (0.865 - 0.929)	0.898 (0.863 - 0.933)	0.895 (0.857 - 0.927)	0.889 (0.851 - 0.920)	0.868 (0.818 - 0.902)



**Figure 1.** Forest plot that displays the AUC values of the methods.

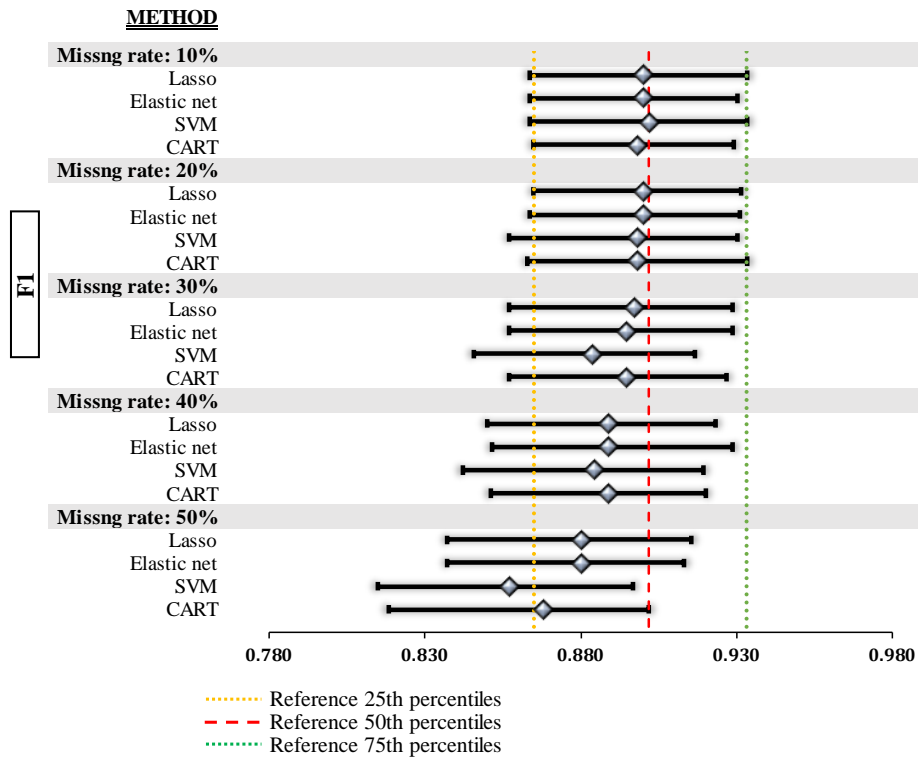


Figure 2. Forest plot that displays the F1 values of the methods.

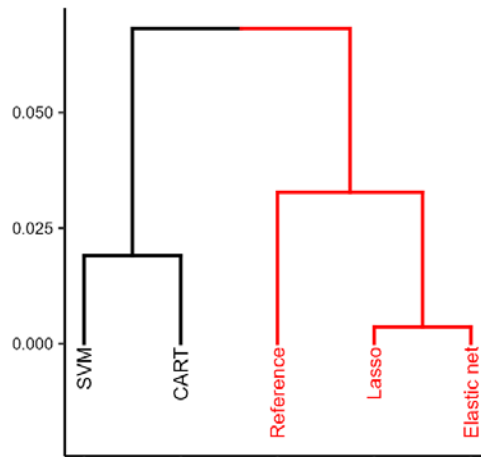


Figure 3. Dendrogram showing the relationship among the imputation models by AUC and F1 values.

#### 4. Discussion

As the field of health continues to evolve, it is anticipated that accurate estimation of missing data will become even more critical to prevent information loss. As high dimensional data containing numerous patient details continue to increase, the incidence of missing data is expected to rise. Therefore, it will be essential to develop techniques to estimate missing data with minimum error to ensure that the models created with this data are as accurate and effective as possible. Schafer and Graham (2002), indicated in their study with real data that if missing observations are deleted from the data set, the statistical power decreases and erroneous inferences are obtained, especially as the missing rate increases.

Therefore, it is recommended to use appropriate methods to handle missing data rather than simply deleting them from the dataset. Liu and De (2015), suggested that the foremost consideration in building an imputation model is to ensure compatibility, failing which, efforts should be directed towards enhancing the predictive accuracy of the imputation model.

The increase in the missing rate directly affects the effectiveness of imputation techniques. The higher the missing rate, the more challenging it becomes to impute the missing values accurately. Qin et al. (2007), showed that an increase in the rate of missing data decreased the accuracy of estimating missing values. Choudhury and Pal (2019), reported that the rise in the rate of missing data

has a detrimental impact on the effectiveness of imputation methods on classification performance, as also found in our study. Furthermore, in our study, although the change in the number of observations and the increase in dimensionality had some impact on the performance of the methods, it was observed that they did not alter the overall performance ranking. Therefore, fixed values for the number of observations and variables were used.

As known, choosing a proper method plays a crucial role dealing with missing data process. Since data structures vary in each dataset, the answer to the question “which method for which dataset” changes. Slade and Naylor (2020), conducted a comparison of the parametric and tree-based imputation methods in the MICE package of the R program. They performed this comparison on a simulated dataset for the regression problem. Their analysis revealed that both parametric and tree-based methods had similar error values and performance. However, the random forests method, which is one of tree-based methods, had the narrowest confidence interval as compared to the other methods. Lavanya et al. (2019), indicated that the lasso imputation method is a highly effective approach to address the challenges associated with missing data in high dimensional datasets. Peña et al. (2019), demonstrated using a real data set that imputation methods such as ridge and lasso, which are based on regularized regression, can estimate missing values with a very low error.

There are only a limited number of studies in the literature that examine the effect of missing data imputation methods on classification performance with different performance evaluation criteria. Liu et al. (2020), conducted a research study analyzing the impact of missing rate on the accuracy of classification. The study revealed that as the rate of missing data increases, the rate of correct classification decreases. Acuna and Rodriguez (2004), found that the imputation method did not significantly affect classification accuracy. However, in their studies, they only used basic and simple imputation methods and worked with datasets that had relatively small amounts of missing data (i.e., between 1% and 20%). Farhangfar et al. (2008), evaluated the effect of some tree-based imputation methods on classification performance on missing data sets with missing rates ranging from 5% to 50%. According to their report, when dealing with data with over 10% missing data, imputation methods tend to improve classification error more than simply ignoring the missing data. However, there is no universally accepted method which could be considered as the best. In our study, we examined the effects of imputation methods on classification performance in high dimensional data, unlike previous literature. Our study revealed that, at lower missing rates, there wasn't a notable variation in performance between the methods in terms of their impact on classification performance. However, as the missing rate increased, tree-based methods were observed to be less effective as compared

to the lasso, and elastic net methods that are based on regularized regression. It was noted that these regularized regression methods performed better than tree-based methods as the missing rate increased.

### 5. Conclusion

In this study, the impact of various imputation methods on classification performance in high dimensional data was evaluated. Our simulation results indicate that regularized regression methods outperform tree-based methods in improving classification on high dimensional data. In the field of data analysis, there are various techniques employed to address the issue of missing data. However, certain methods may not be efficient in handling high dimensional data due to their underlying theoretical framework. As a result, it is crucial to carefully select the appropriate method to avoid information loss that is common in high dimensional data and to enhance the accuracy of predictive models.

### Author Contributions

The percentages of the authors' contributions are presented below. The authors reviewed and approved the final version of the manuscript.

	B.V.	İ.K.Ö.	M.T.
C	35	35	30
D	40	40	30
S	10	45	45
DCP	50	25	25
DAI	30	50	20
L	25	25	50
W	60	20	20
CR	40	30	30
SR	45	45	10
PM	40	40	20
FA	50	30	20

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision, PM= project management, FA= funding acquisition.

### Conflict of Interest

The authors declared that there is no conflict of interest.

### Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans.

### Acknowledgements

I would like to thank the anonymous reviewers and editors for their valuable comments and suggestions regarding this article.

References

- Acuna E, Rodriguez C. 2004. The treatment of missing values and its effect on classifier accuracy. In: Classification, Clustering, and Data Mining Applications: Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, July 15–18, Chicago, USA, pp: 639-647.
- Breiman L. 1995. Better subset regression using the nonnegative garrote. *Technometrics*, 37(4): 373-384.
- Breiman L. 2017. Classification and regression trees. Routledge, New York, USA, 1st ed., pp: 368.
- Chang LY, Chen WC. 2005. Data mining of tree-based models to analyze freeway accident frequency. *J Saf Res*, 36(4): 365-375.
- Choudhury SJ, Pal NR. 2019. Imputation of missing data with neural networks for classification. *Knowledge-Based Syst*, 182: 104838.
- Clark LA, Pregibon D. 2017. Tree-based models. In: Hastie T, Chambers J, editors. *Statistical models in S*, Routledge, Oxfordshire, UK, pp: 377-419.
- Cortes C, Vapnik V. 1995. Support-vector networks. *Mach Learn*, 20: 273-297.
- Deng Y, Chang C, Ido MS, Long Q. 2016. Multiple imputation for general missing data patterns in the presence of high-dimensional. *Data Sci Rep*, 621689, 6(1): 21689.
- Elith J, Leathwick JR, Hastie T. 2008. A working guide to boosted regression trees. *J Anim Ecol*, 77(4): 802-813.
- Enders CK. 2022. Applied missing data analysis. Guilford Press, New York, USA, 2nd ed., pp: 546.
- Farhangfar A, Kurgan L, Dy J. 2008. Impact of imputation of missing values on classification error for discrete data. *Pattern Recognit*, 41(12): 3692-3705.
- Fawcett T. 2006. An introduction to ROC analysis. *Pattern Recognit Lett*, 27(8): 861-874.
- Friedman J, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J Stat Software*, 33(1): 1-22.
- Hanley JA, McNeil BJ. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1): 29-36.
- Hastie T, Tibshirani R, Friedman JH, Friedman JH. 2009. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, USA, 2nd ed., pp: 737.
- Jadhav A, Pramod D, Ramanathan K. 2019. Comparison of performance of data imputation methods for numeric dataset. *Appl Artif Intell*, 33(10): 913-933.
- Lavanya K, Reddy L, Eswara Reddy B. 2019. A study of high-dimensional data imputation using additive LASSO regression model. In: Behera HS, Nayak J, Naik B, Abraham A, editors. *Computational intelligence in data mining*. Springer, Singapore, pp: 19-30.
- Little RJ, Rubin DB. 2019. Statistical analysis with missing data. John Wiley & Sons, New York, USA, 3rd ed., pp: 449.
- Liu CH, Tsai CF, Sue KL, Huang MW. 2020. The feature selection effect on missing value imputation of medical datasets. *Appl Sci*, 10(7): 2344.
- Liu Y, De A. 2015. Multiple imputation by fully conditional specification for dealing with missing data in a large epidemiologic study. *Int J Stat Med Res*, 4(3): 287-295.
- Loh WY. 2011. Classification and regression trees. *Interdiscip Rev Data Min Knowl Discov*, 1(1): 14-23.
- Nawar S, Mouazen AM. 2017. Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line Vis-NIR spectroscopy measurements of soil total nitrogen and total carbon. *Sensors*, 17(10): 2428.
- Patil AR, Kim S. 2020. Combination of ensembles of regularized regression models with resampling-based lasso feature selection in high dimensional data. *Mathematics*, 8(1): 110.
- Peña M, Ortega P, Orellana M. 2019. A novel imputation method for missing values in air pollutant time series data. In: 2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI), November 11-15, Guayaquil, Ecuador, pp: 1-6.
- Przednowek K, Wiktorowicz K. 2013. Prediction of the result in race walking using regularized regression models. *J Theor Appl Comput Sci*, 7(2): 45-58.
- Qin Y, Zhang S, Zhu X, Zhang J, Zhang C. 2007. Semi-parametric optimization for missing data imputation. *Appl Intell*, 27(1): 79-88.
- Rubin DB. 1988. An overview of multiple imputation. *Proc Surv Res methods Sect Am Stat Assoc*, 16: 79-84.
- Schafer JL, Graham JW. 2002. Missing data: our view of the state of the art. *Psychol methods*, 7(2): 147-177.
- Schapire RE. 2003. The boosting approach to machine learning: An overview. In: Denison DD, Hansen MH, Holmes CC, Mallick M, Yu B, editors. *Nonlinear estimation and classification*. Springer, New York, 2023rd ed., pp: 149-171.
- Slade E, Naylor MG. 2020. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Stat Med*, 39(8): 1156-1166.
- Stekhoven DJ, Bühlmann P. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1): 112-118.
- Tharwat A. 2021. Classification assessment methods. *Appl Comput Inform*, 17(1): 168-192.
- Tian Z, Xiao J, Feng H, Wei Y. 2020. Credit risk assessment based on gradient boosting decision tree. *Procedia Comput Sci*, 174: 150-160.
- Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*, 58(1): 267-288.
- Yin X, Levy D, Willinger C, Adourian A, Larson MG. 2016. Multiple imputation and analysis for high-dimensional incomplete proteomics data. *Stat Med*, 35(8): 1315-1326.
- Zhang S, Gong L, Zeng Q, Li W, Xiao F, Lei J. 2021. Imputation of gps coordinate time series using missforest. *Remote Sens*, 13(12): 2312.
- Zhang Z. 2016. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Ann Transl Med*, 4(2): 30.
- Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevskaya O. 2019. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med*, 7(7): 152.
- Zhao Y, Long Q. 2016. Multiple imputation in the presence of high-dimensional data. *Stat Methods Med Res*, 25(5): 2021-2035.
- Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*, 67(2): 301-320.