The North African Journal of Scientific Publishing (NAJSP)
مجلة شمال إفريقيا للنشر العلمي (NAJSP)
E-ISSN: 2959-4820
Volume 4, Issue 4, 2026
Page No: 166-180
Website: https://najsp.com/index.php/home/index
SJIFactor 2024: 5.49 معامل التأثير العربي (AIF) 2025: 0.69 ISI 2024: 0.696

A Comparative Evaluation of Imputation Methods for Support Vector Regression under Missing at Random

Bassma Ali Altmtam[1], Alsaidi M. Altaher[2*]
[1,2]Department of Statistics, Faculty of Sciences, Sebha University, Sebha, Libya

تقييم مقارن لطرق التعبئة البديلة لانحدار المتجهات الداعمة في ظل الفقدان العشوائي للبيانات

بسمه علي أحمد التمتام[1]، السعيدي م. الطاهر[2*]
[1،2]قسم الإحصاء، كلية العلوم، جامعة سبها، سبها، ليبيا

*Corresponding author: als.altaher@sebhau.edu.ly

Received: August 06, 2026	Accepted: October 16, 2026	Published: October 30, 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract:
Handling missing data remains a fundamental challenge in the context of Support Vector Regression (SVR). This study aims to compare the performance of five missing data imputation methods (Mean, Median, KNN, MICE, MissForest) and their impact on SVR accuracy, These methods range from single-value imputation (Mean, Median) to similarity-based (KNN) and relationship-based (MICE, MissForest) approaches. The study employed these methods through two analytical tracks using a Monte Carlo simulation study, which varied the noise level, number of predictors, and sample size, alongside the real-world Boston Housing Dataset with missingness rates of 5%, 10%, 20%, 30% applied in both tracks, and performance was evaluated using RMSE, MAE, and R^2. Results showed that increasing missingness negatively affects SVR accuracy in both analytical tracks. MICE and MissForest consistently outperformed the traditional methods: mean and median imputation were the weakest across every simulation condition and the real-data application, while KNN performed competitively, even outperforming MICE on the real data at higher missingness rates. MissForest excelled at low-to-moderate noise levels and MICE at high noise. Results showed also that increasing sample size improves the performance for all methods.

Keywords: Missing at Random (MAR), Imputation, Support Vector Regression, Simulation, Boston Housing Dataset.

الملخص:
تُعد معالجة البيانات المفقودة تحديًا رئيسيًا في انحدار متجهات الدعم (SVR). تهدف هذه الدراسة إلى مقارنة أداء خمس طرق لتعويض البيانات المفقودة، وهي: المتوسط، الوسيط، k-NN، MICE، MissForest، وتأثيرها في دقة نموذج SVR.تتراوح هذه الطرق بين التعويض بقيمة مفردة (المتوسط والوسيط) ، والتعويض القائم على التشابه (k-NN) ، والتعويض القائم على العلاقات بين المتغيرات (MissForest , MICE).وظفت الدراسة هذه الطرق باستخدام مسارين تحليليين من خلال دراسة محاكاة مونت كارلو حكمت فيها مستويات الضوضاء، وعدد المتغيرات التفسيرية وحجم العينة،

إلى جانب بيانات Boston Housing الحقيقية. وذلك عند نسب فقد بلغت 5%، 10%، 20%، و30% في كلا المسارين. وقيس الأداء باستخدام مقاييسRMSE، MAE ، R^2. أظهرت النتائج أن زيادة نسبة الفقد تؤثر سلبًا في دقة نموذجSVR في كلا المسارين التحليليين وتفوقت طريقتا MICE وMissForest، باستمرار على الطرق التقليدية: المتوسط والوسيط كانا الأضعف في جميع حالات المحاكاة والبيانات الحقيقية، بينما أدت KNN أداءً تنافسياً بل تفوقت على MICEفي البيانات الحقيقية عند نسب الفقد الأعلى. وتفوق طريقة MissForestعند مستويات الضوضاء المنخفضة والمتوسطة، بينما تفوقت MICE عند مستوى الضوضاء المرتفع . النتائج بينت ايضا ام زيادة حجم العيمة يحسن دقة التقدير لكل الطرق.

الكلمات المفتاحية: البيانات المفقودة عشوائيا، تعويض البيانات، الانحدار بمتجهات الدعم، المحاكاة، بيانات بوسطن للإسكان.

Introduction:

Support Vector Regression (SVR) is a machine learning algorithm which has been developed in the past few decades for regression problems. Its advantage over the traditional regression methods is due to the fact that it is more general and efficient in handling nonlinear data. SVR can accurately predict as it is able to detect complex patterns. In real world data sets, however, missing values can be found and these can lower the predictive accuracy and increase the bias. When imputing an SVR model the model may not be unbiased and may not capture the structure of the data. The model may therefore be affected by missing data which impacts its generalization capability. Given this practical context, Batra et al. [1] built an ensemble imputation model which involved three traditional imputations techniques: mean, K-nearest neighbors (K-NN), and Iterative Imputation. They evaluated it with real health care data sets with different degrees of missingness using SVR along with two regressors, XGBoost and Random Forest Regressor. The ensemble was found to be more effective than any of the three imputation methods that were used individually. However, the study left out advanced techniques of missing data imputation, and only employed real world data; no simulation experiments were used to support the results. In the study on development of non-linear predictive model by Ibrahim and Wibowo [2], they used four kinds of kernel functions of the SVR: linear, sigmoid, polynomial and radial basis function (RBF). In addition, they tried two missing value imputation approaches: Mean Imputation and ordinary linear regression (OLR). Among all the methods, Mean Imputation had the lowest MSECV value and RBF kernel was the most suitable performance. The study did not, however, report which type of missingness was investigated and how it was found out because this may impact the performance of the imputation later on. Additionally, it did not report on the percentage/magnitude of missingness, making it hard to tell what missing-data issue it addressed. Meyer et al. [3] set out to predict the percentage of women in science and technology from a set of independent variables with the hope that the SVR would be able to capture nonlinear relationships that could not be captured by the simpler models. They experimented with two types of kernels: linear and radial basis function (RBF). Among the five regression models studied in this paper, Ridge, Linear, ElasticNet, Lasso regression and SVR, the latter gave the best score in the cross validation (CV score = 0.7697) and the lowest prediction error (MAE = 0.2677). The RBF kernel was a more suitable kernel for the data. Many imputation methods, such as KNN, mean/median and MICE are mentioned in the paper but only KNN was applied. Nor was there an element of simulation to the study, which could have provided further support for the empirical results. Shams and Elshewey [4] applied Support Vector Regression (SVR) to predict the Water Quality Index (WQI), replacing missing values with the feature mean through mean imputation and using seven chemical water features. The model achieved strong predictive performance (R^2 = 99.1%), ranking second behind MLP and ahead of KNN and DT. Still, the result raises the possibility of overfitting, which could limit the model's ability to generalize to new data. This concern is especially relevant because the study did not report the proportions of missing values. El-Badisy et al. [5] conducted a Monte Carlo simulation based on 500 survival-time datasets generated from a Weibull distribution. They additionally analyzed real data from a breast cancer survival study in Morocco, introducing 30% missing values under the Missing at Random (MAR) mechanism and using the Cox proportional hazards model for estimation. Among the methods examined, the two single-

imputation (SI) approaches, MissMDA and MissForest, performed better than the others. The study was limited, though, to one setting: a 30% missingness rate under a single MAR mechanism, evaluated over 500 simulation replications. Most earlier studies have important limitations. They generally did not compare traditional and advanced missing-data imputation methods within the same analysis. Their evaluations also relied largely on real datasets, without simulation designs that would permit methodological control or systematic changes in sample size, number of variables, and noise level. In several cases, the missingness mechanism and proportion of missing data were not reported. Only a small number used SVR as the final predictive model after applying different imputation methods. This paper addresses the gap by applying SVR after missing data have been handled, then comparing three traditional approaches-mean imputation, median imputation, and k-nearest neighbors (KNN)-against two advanced ones: MICE and MissForest. It examines the methods along two analytical tracks, systematically varying both the missingness rate and sample size. One uses real data from the Boston Housing dataset; the other relies on carefully designed, controlled simulations, allowing the findings to be generalized across multiple data contexts.

The rest of this paper is organized as follows: Section 2 covers the material and methods; Section 3 reports the experimental results; and Section 4 offers the conclusions.

Material and Methods:

Support Vector Regression (SVR):

Support Vector Regression (SVR) is an extension of Support Vector Machines (SVM) that was introduced for regression by Drucker and Vapnik [6]. The SVR can model nonlinear relationships among variables and is relatively resistant to overfitting. The aim is to find a predictive function $f(x)$ such that its deviation from the observed values yi stays within a margin ε (epsilon), while keeping the function as flat as possible.

Typically, the SVR model starts with the linear case, and then is extended to nonlinear case by using different kernel functions. Its regression function is written as**:**

Where w is the weight vector, and b is the bias term. The desired flatness is achieved by formulating the problem as a convex optimization problem:

subject to:

$$y_i - \langle w, x_i \rangle - b \leq \varepsilon$$
$$\langle w, x_i \rangle + b - y_i \leq \varepsilon$$

Since these constraints might not hold for every data point, slack variables ξi and ξi∗ are added to permit deviations exceeding ε. The optimization problem then takes **the following form:**

$$min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i *) \qquad (2)$$

subject to:

$$y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i$$
$$\langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i *$$
$$\xi_i, \xi_i * \geq 0$$

where the positive constant C acts as a regularization parameter, balancing the flatness of f with the allowed size of deviations beyond ε. The input vector x is replaced by a mapping ϕ(x) into a higher-dimensional feature space, allowing the model to represent nonlinear relationships among variables and transforming the regression function into:

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i *)K(x_i, x) + b \qquad (3)$$

By introducing the Lagrange multipliers αᵢ and αᵢ*, the inner product can be expressed through a kernel function, yielding the final SVR prediction function:

For this study, the Gaussian Radial Basis Function (RBF) kernel was selected as it can capture the complex nonlinear relationships. It is defined as:

$$K(x_i, x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right) \qquad (4)$$

where γ > 0 is a parameter that controls the range of influence of each training observation on neighboring observations [7].

Missing-Data Mechanisms:

Understanding why values go missing matters because recognizing the underlying mechanism helps researchers choose suitable ways to address the problem, improving the validity and accuracy of their findings. According to Rubin [8] and Enders [9], missing-data mechanisms are classified into three main types:

1. **Missing Completely at Random (MCAR):** With MCAR, missing data occur entirely at random, without being affected by observed or unobserved values. Specifically:

$$p(M \mid Y_{obs}, Y_{mis}) = p(M) \qquad (5)$$

where M denotes the missingness matrix, and Y_{obs} and Y_{mis} refer to the observed and missing variables in the dataset, respectively.

2. **Missing at Random (MAR):** With MAR, the likelihood that a value is missing depends only on the observed values and not on the missing values themselves:

$$p(M \mid Y_{obs}, Y_{mis}) = p(M \mid Y_{obs}) \quad (6)$$

3. **Missing Not at Random (MNAR):** With MNAR data, whether a value is missing depends partly on the value that is missing itself, even after the observed variables are taken into account.

$$p(M \mid Y_{obs}, Y_{mis}) = p(M \mid Y_{mis}) \quad (7)$$

For both analytical tracks, missing values were introduced under the MAR assumption. MAR was chosen for three main reasons: (1) it is more realistic than MCAR in most practical settings, as missingness often depends on other observed variables rather than occurring purely at random; (2) it allows advanced imputation methods to exploit the correlations among observed variables, thereby improving imputation accuracy; and (3) it provides a fair and rigorous basis for comparing imputation methods, since the missingness mechanism is neither trivially random (MCAR) nor intractably non-random (MNAR).

Missing-Data Imputation Methods:

Single or simple imputation techniques, including mean, median and the k-nearest neighbors (k-NN) method (falls between these basic techniques and more advanced imputation approaches), are usually easy to understand and apply, especially when only a very small number of the observation is missing. Their practical ease, however, does not eliminate several limitations. Some ignore relationships between variables: mean imputation, for instance, fills in missing values with the average of the observed values without considering potential associations between the affected variable and the others. They can also change the distribution of the original data, particularly if the distribution is distorted or there are a lot of missing values [10].

Those restrictions might end up reducing the ability of SVR to predict the results in practice, since SVR is based on the ability to accurately model the nonlinear relationship between variables. This can be done more flexibly with more sophisticated imputation methods but the outcome may vary widely depending on the imputation model used. Consequently, when using SVR in this study, the two more advanced missing value imputation methods Multiple Imputation by Chained Equations (MICE) and MissForest Imputation were also compared to the conventional methods.

1. **Multiple Imputation by Chained Equations (MICE):**

As a flexible approach to imputing missing values, Van Buuren and Groothuis-Oudshoorn [11] proposed to view MICE. It creates several full data sets in a series of conditional regression models, one for each variable with missing data. The resulting statistical estimates are closer to capturing uncertainty added in the imputation process.

Steps in the MICE Algorithm:

- **Step 1:** Use a simple guess at each of the missing values (mean or median). This results in a "full" data set at first.
- **Step 2:** Obtain an appropriate conditional regression model for the variable with missing values, where the missing values of the variable are the dependent variables in the model and the model is estimated from the known values of the dependent variable only.

- **Step 3:** For every variable with missing values, repeat Steps 1 and 2. Use the fitted model associated with that variable to predict the missing entries and replace them.
- **Step 4:** Repeat the full imputation procedure to create m complete datasets. Then apply the SVR model to each one.

2. MissForest Algorithm:

MissForest is a nonparametric imputation method for missing values proposed by Stekhoven and Bühlmann[12]. It predicts missing entries in each variable, based on the relationships the variable has with the other variables, using Random Forests. This renders MissForest particularly suitable for data with complex and nonlinear associations and comparative studies have shown it to be superior to several different imputation procedures.

Steps of the MissForest Algorithm:

- **Step 1:** The first step is to begin by simply imputing each missing value (this can be done in a very simple way, such as mean or median imputation). Then list the variables in the order of the smallest to the largest percentage of missing values.
- **Step 2:** Fit a Random Forest model using as the dependent variable the first variable in the ordered sequence. The model is fitted on the observed values of the variable and the missing values are predicted and filled with the model.
- **Step 3:** Go to the next variable and repeat steps 2 and 3 with the previously imputed values of the variables. Do this until all variables have been imputed once; until you have completed an iteration.
- **Step 4:** Go through all of the iterative cycles from steps 2 and 3. Once this is done, the algorithm returns a new data set, the SVR model is directly applied to the single data set.

Model Evaluation Metrics:

Three evaluation criteria were used to compare the imputation methods and evaluate model performance.

1. **Root Mean Square Error** $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (8)$$

RMSE is a statistical measure used to evaluate predictive accuracy. Lower RMSE values indicate greater prediction accuracy and better model performance.

2. **Mean Absolute Error** $$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (9)$$

MAE represents the average absolute difference between the actual and predicted values. Lower values indicate smaller prediction errors.

3. **Coefficient of Determination** $$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (10)$$

The coefficient of determination measures the proportion of variation in the response variable that the model explains.

Results and Discussions:

Experimental Design:

The study compared the five imputation methods along two analytical tracks, using both simulated data and real Boston Housing data. For each track, the data were split into a 70% training set and a 30% testing set. Missing values were subsequently added to each set independently at predefined rates, assuming MAR The missing values were replaced separately with each imputation method. The RBF-SVR model's predictive performance was then assessed on the resulting datasets using the evaluation metrics defined earlier, with the model parameters tuned via Grid Search for $C \in \{0.1, 10\}$ $\gamma \in \{0.01, 0.1\}$, $\varepsilon=0.1$ using 3-fold Cross-Validation.

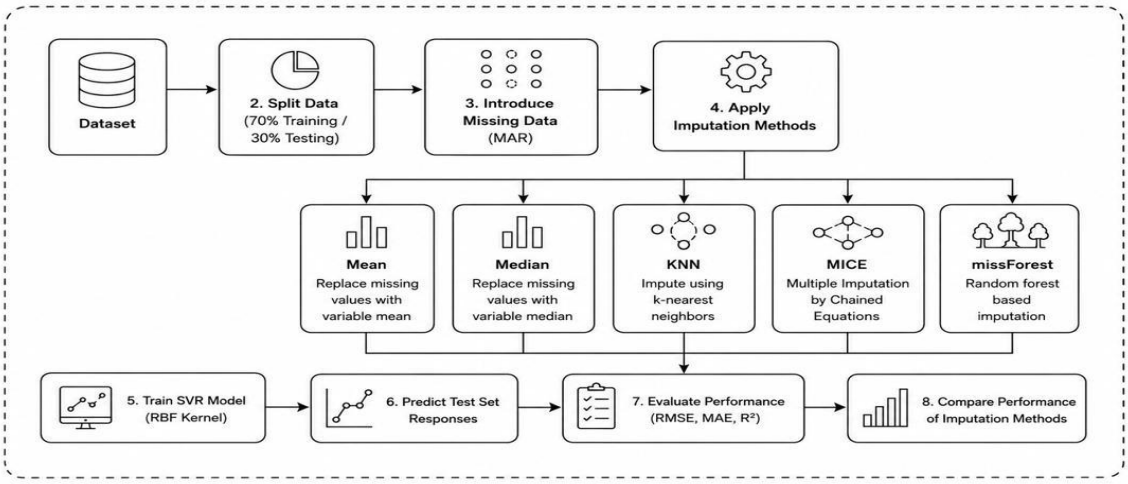


Figure (1): summarizes the experimental design of the study
Source: prepared by the researchers based on the study procedures

Track One: Simulation Experiments:

The study generated simulated data by defining the response variable (y) as a linear function of the independent variables plus a noise term. This setup made it possible to assess the imputation methods under controlled conditions, with the characteristics of those conditions varied systematically, the

independent variables first generated from multivariate normal distribution $X_i \sim \mathrm{MVN}(0, \Sigma)$ where Σ is the variance-covariance matrix with a correlation coefficient (ρ=0.6) between each pair of variables-was necessary to give the relational imputation methods(K-NN, MICE, and Missforest) a genuine opportunity to exploit the actual correlations among the variables. Across all simulation settings, the

intercept remained fixed at β_0=1 for all configurations, and β_i were randomly generated anew at each

replication from a Uniform distribution $\beta_i \sim U[-2, 2]$ giving equal probability to every value between -2 and 2 thus covering across replications the full range of relationship strength(from weak to strong, positive to negative) at every one of the 500 simulation replications ,according to the following model:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \varepsilon ; i = 1,2,\ldots,p \quad \varepsilon \sim N(0, \sigma^2)$ where ε denotes the noise term. The design varied three sample sizes (N=50,100,200), three model dimensions (P=2,3,5), and missingness rates of 5%,10%, 20%,30%. It also included three noise levels: σ=0.5 for low noise, σ=1 for moderate noise, and σ=2 for high noise.

Track Two: Boston Housing Data:

The Boston Housing dataset was obtained from the standard "MASS" package and served as the real-world dataset for the study [13]. It contains 506 observations, 13 continuous explanatory with the

exclusion of variable B; $B = 1000(Bk - 0.63)^2$ where Bk is the proportion of Black residents in the town (due to its reliance on a racial assumption that is scientifically and ethically outdated) and the response variable MEDV that denotes the median value of owner-occupied homes in ($1000s). Before analysis, all variables were standardized. Missing values were subsequently added to the variables at four preset rates: 5%, 10%, 20%, and 30%. The missingness was restricted to four continuous variables, deliberately chosen to cover four different dimensions, as shown in Table (1): The description of Boston Haousing data.

Table (1): The description of Boston Haousing data

Variable	Dimension Represented
RM (Rooms Mean)	Housing structure
AGE (Age of building)	Building age
DIS(Distances)	Location
LSTAT (Lower Status)	Socioeconomic status

Source: [13], with modification by excluding variable (B)

All analyses for this study were carried out in R (version 4.5.2). Missing data were processed with the mice, missForest, VIM, and k-NN packages, and the e107 package was used to implement the SVR model. The analysis also included dplyr and ggplot2.

Main Results:

- **Results of Track One: Simulation Experiments:** This section reports the main findings from the simulated data and assesses how the SVR Performance varies with the noise level, the number of predictors, and missingness rate. Table (2) lists the best Performing imputation method under each condition, and Table (3) reports the main effect of each factor. All reported values are averaged over the 500 replications and three sample sizes (N=50, 100, and 200).

Table (2): Summary of imputation Performance by Noise Level, number of predictors, and missing rate

Noise (σ)	Predictors	Missing Rate	Best Method	RMSE	MAE	R^2
0.5	2	5%	MICE	0.535	0.430	0.796
		10%	MICE	0.537	0.431	0.788
		20%	MissForest	0.544	0.434	0.773
		30%	MissForest	0.565	0.451	0.778
	3	5%	MICE	0.554	0.443	0.856
		10%	MissForest	0.557	0.445	0.856
		20%	MissForest	0.587	0.468	0.839
		30%	MissForest	0.615	0.488	0.818
	5	5%	MissForest	0.600	0.477	0.901
		10%	MissForest	0.628	0.498	0.891
		20%	MissForest	0.668	0.529	0.880
		30%	MissForest	0.731	0.573	0.856
1.0	2	5%	MICE	1.037	0.835	0.553
		10%	MICE	1.043	0.840	0.547
		20%	KNN	1.067	0.858	0.522
		30%	MissForest	1.071	0.860	0.502
	3	5%	MICE	1.066	0.858	0.631
		10%	MICE	1.079	0.868	0.626
		20%	MissForest	1.090	0.877	0.624
		30%	MissForest	1.121	0.901	0.616
	5	5%	MICE	1.118	0.899	0.733
		10%	MICE	1.143	0.917	0.728
		20%	MissForest	1.180	0.942	0.693
		30%	MissForest	1.219	0.977	0.673
2.0	2	5%	MICE	2.070	1.665	0.234
		10%	MICE	2.070	1.666	0.214
		20%	MICE	2.074	1.669	0.226
		30%	MICE	2.102	1.691	0.216
	3	5%	MICE	2.114	1.703	0.294
		10%	MICE	2.091	1.686	0.295
		20%	MICE	2.115	1.704	0.303
		30%	MICE	2.140	1.723	0.291
	5	5%	MICE	2.163	1.740	0.406
		10%	MICE	2.152	1.731	0.412
		20%	MICE	2.199	1.767	0.400
		30%	MissForest	2.251	1.808	0.356

Note: "Best Method"=the imputation method with the lowest mean (RMSE, MAE) in the condition.

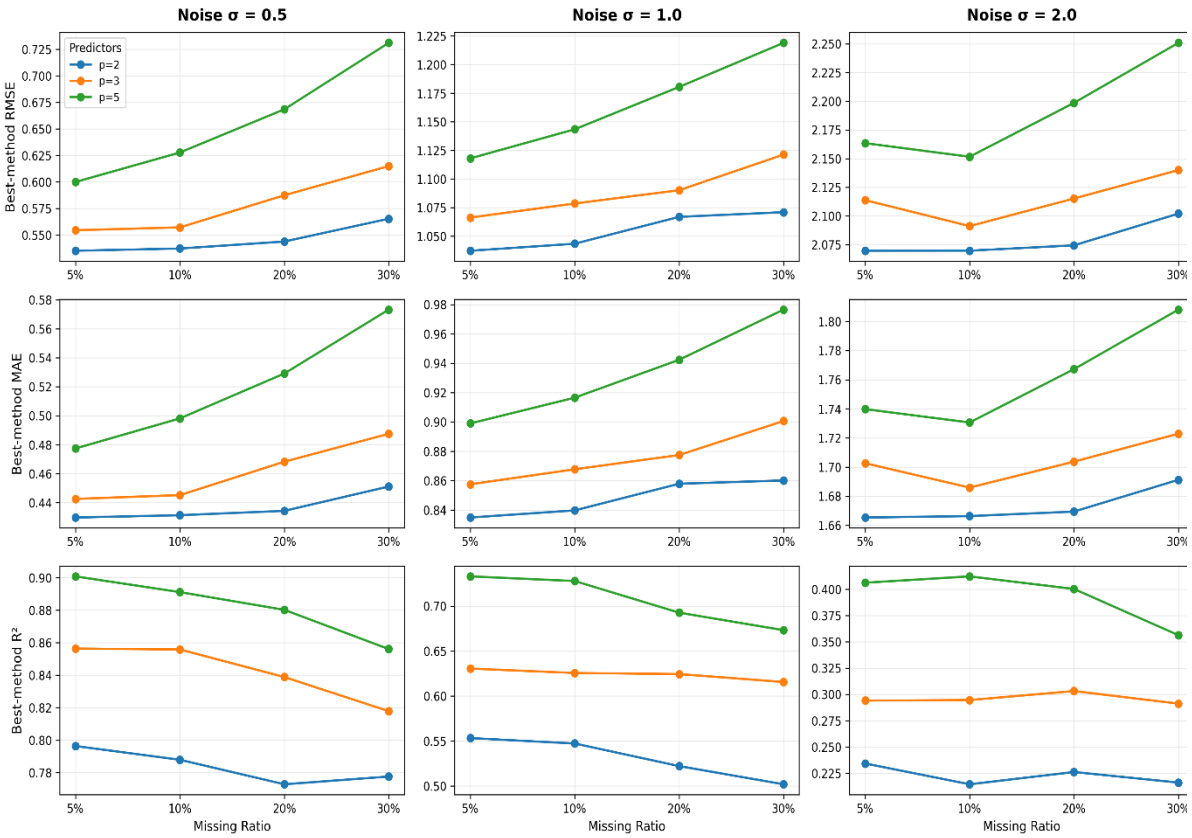


Figure (2): Best method by metrics (RMSE, MAE, and R^2) across noise levels, the number of predictors, and missingness rate

The results show that noise is the most influential factor affecting model accuracy, where under low noise (σ=0.5), MissForest outperforms the other methods in most conditions, especially as the number of predictors increases, while MICE only outperform at the lowest missingness rates. Under moderate noise (σ=1), MICE outperforms at low missingness and MissForest at higher missingness with one exception showing KNN outperforming (p=2 at 20%). Under high noise (σ=2), the differences between methods are small, and MICE outperforms in most conditions. Overall, the best method's error increases with higher missingness and more predictors, except for a slight fluctuation under high noise. Figure (2) illustrates these patterns.

Table (3): Main effects of Noise Level, number of predictors, and missing rate

Factor	Level	Avg RMSE	Avg MAE	Avg R^2	Best Method
Noise (σ)	0.5	0.625	0.496	0.823	MissForest
	1.0	1.125	0.903	0.608	MissForest
	2.0	2.153	1.732	0.286	MICE
Predictors	2	1.238	0.995	0.503	MICE
	3	1.284	1.031	0.573	MissForest
	5	1.381	1.105	0.641	MissForest
Missing Ratio	5%	1.261	1.013	0.593	MICE
	10%	1.272	1.022	0.585	MICE
	20%	1.311	1.051	0.568	MissForest
	30%	1.360	1.089	0.544	MissForest

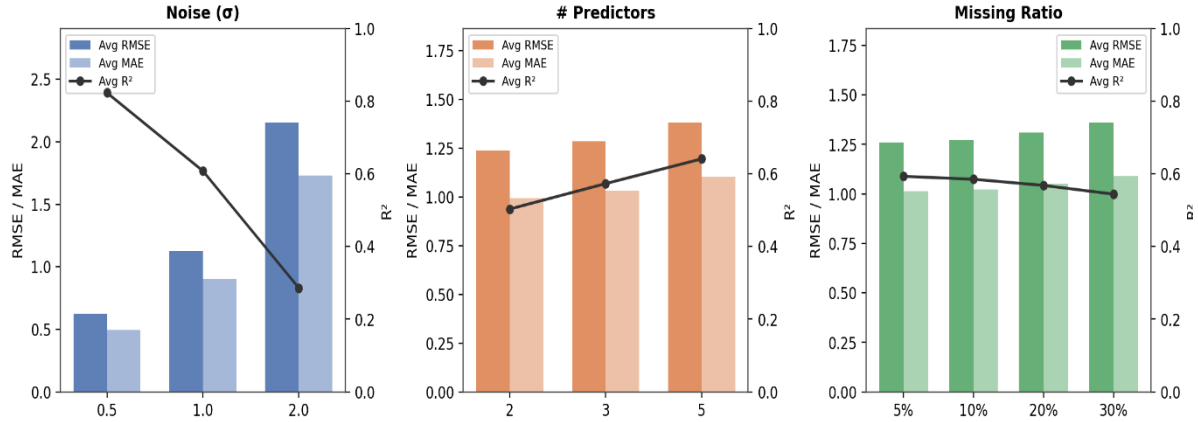


Figure (3): Main effects of Noise Level, number of predictors, and missing rate on (RMSE, MAE, and R^2)

"Best method" here reflects the overall average and may differ from the best method in each individual condition shown in table (2). Noise is also the most influential factor, as RMSE rises from 0.625 to 2.153 and R^2 falls from 0.823 to 0.286 between σ=0.5 and σ= 2. Predictors and missingness rates have a much smaller effect: more Predictors raise both RMSE and R^2, because the additional Predictors also increase the total variance in the dependent variable, so the model explains a larger proportion of it even as the absolute error), while more missingness raises RMSE and lowers R^2. MissForest performs best under low-to moderate noise and with more predictors and missingness, while MICE performs best under high noise and fewer Predictors and lower missingness. Figure (3) shows noise as the dominant factor across RMSE, MAE, and R^2, with smaller, more gradual effects from the other two factors.

Table (4): Mean (RMSE, MAE, and R^2) of SVR by sample size and imputation method

N	Metric	Mean	Median	KNN	MICE	MissForest	Overall Mean
50	RMSE	1.394	1.392	1.364	**1.346**	1.353	1.370
	MAE	1.125	1.124	1.101	**1.088**	1.092	1.106
	R^2	0.480	0.481	0.498	**0.517**	0.504	0.496
100	RMSE	1.311	1.312	1.275	1.271	**1.266**	1.287
	MAE	1.049	1.049	1.021	1.018	**1.014**	1.030
	R^2	0.578	0.578	0.596	**0.600**	0.599	0.590
200	RMSE	1.272	1.273	1.230	1.234	**1.223**	1.247
	MAE	1.014	1.014	0.983	0.986	**0.977**	0.995
	R^2	0.620	0.620	0.638	0.637	**0.640**	0.631

Note: values are averaged over the 500 replications, Noise Level, number of predictors, and missing rate on (RMSE, MAE, and R^2). Bold indicates the best method for each metric.

Table (4) shows that sample size improved performance for all methods: RMSE fell from 1.370 at N=50 to 1.247 at N=200 and R^2 rose from 0.496 to 0.631. MICE was best at N=50 and MissForest at N ≥100, with small differences among the top methods. Mean and median imputation were the weakest at every N. Figure (4) illustrates these trends.

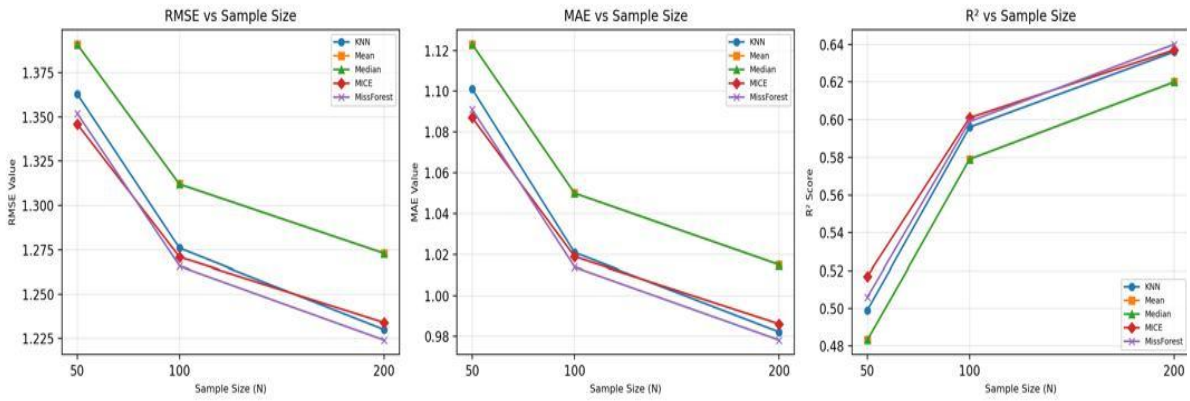


Figure (4): Effect of sample size on (RMSE, MAE, and R^2) for the imputation methods

- **Results from Track Two: Real Data (Boston Housing Data):**

This section reports the results of applying the imputation methods to the Boston Housing dataset at missingness rates of 5%, 10%, 20%, and 30%, along with how each method affected SVR performance after imputation.

Table (5): Mean RMSE of SVR Performance for Boston Housing Data

Missing Rate	Mean	Median	KNN	MICE	MissForest	Overall Mean
5%	3.856	3.860	3.713	3.632	3.701	3.752
10%	3.978	4.007	3.792	3.744	3.772	3.859
20%	4.038	4.116	3.826	3.844	3.799	3.925
30%	4.184	4.260	3.921	4.042	3.894	4.060
Overall Mean	4.014	4.061	3.813	3.815	3.792	3.899

Table (4) shows the error increased as missing- data rate rose, since the SVR-RBF builds its prediction function based on distances between data points via RBF kernel, so any distortion in the values directly affects its accuracy. MissForest had the lowest overall mean error, at 3.792, followed by KNN at 3.813 and then MICE (3.815). Median imputation produced the highest overall mean RMSE, 4.061.

Table (6): Mean MAE of SVR Performance for Boston Housing Data

Missing Rate	Mean	Median	KNN	MICE	MissForest	Overall Mean
5%	2.354	2.354	2.310	2.240	2.305	2.313
10%	2.404	2.408	2.346	2.274	2.336	2.354
20%	2.476	2.507	2.392	2.343	2.367	2.417
30%	2.579	2.607	2.449	2.443	2.414	2.499
Overall Mean	2.453	2.469	2.375	2.325	2.355	2.395

Table (5) reports the mean absolute error results. The average MAE for the different methods increased from 2.313 at a 5% missingness rate to 2.499 at 30%. MICE and MissForest were the top overall performance with 2.325 and 2.355 respectively. The highest mean absolute error was obtained by the median imputation method, being equal to 2.469.

Table (7): Mean Coefficient of Determination (R^2) of SVR Performance Boston Housing Data

Missing Rate	Mean	Median	KNN	MICE	MissForest	Overall Mean
5%	0.819	0.818	0.832	0.839	0.833	0.828
10%	0.807	0.804	0.825	0.828	0.827	0.818
20%	0.800	0.792	0.821	0.819	0.823	0.811
30%	0.788	0.780	0.814	0.803	0.817	0.800
Overall Mean	0.803	0.799	0.823	0.822	0.825	0.814

The decline in R^2 reflects SVR's diminishing ability to explain variance. The SVR model was able to account for less of the variation in the response variable as data became missing. SVR with MissForest imputation method has the best mean R^2 value of 0.825. And KNN, MICE had the closest value to it, 0.823, 0.822 respectively, Median imputation had the lowest value of 0.799.

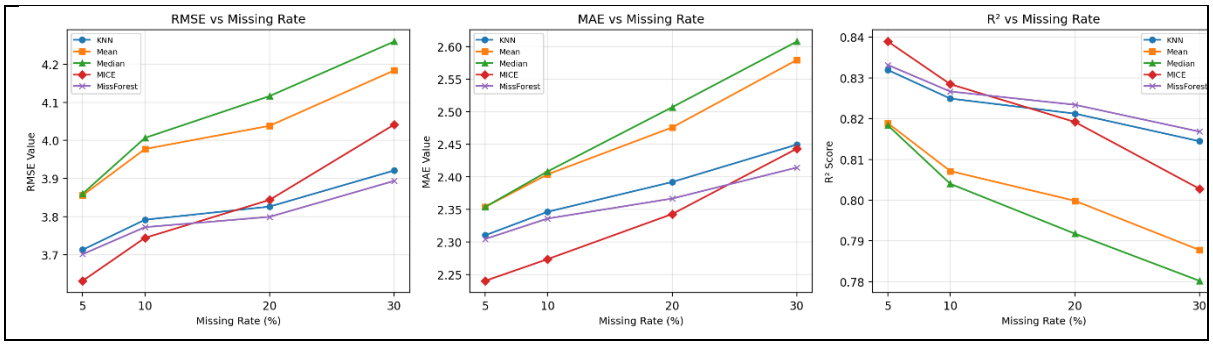


Figure (5): Evaluation Metrics (RMSE, MAE, and R2) for Five Imputation Methods across Missingness Rates Boston Housing Data

Table (8): Optimal SVR Hyperparameters (C, γ) across imputation method-Monte carlo Simulation

Imputation Method	Mean SVR Hyperparameters	
	Best C	Best gamma(γ)
Mean	9.5145	0.0234
Median	9.5242	0.0234
KNN	9.5580	0.0212
MICE	9.4738	0.0224
MissForest	9.5877	0.0208

Table (9): Optimal SVR Hyperparameters (C, γ) across imputation methods-Boston Housing Data

Imputation Method	Mean SVR Hyperparameters	
	best_ C	best_ gamma(γ)
Mean	10.0	0.0270
Median	10.0	0.0283
KNN	10.0	0.0387
MICE	10.0	0.0334
MissForest	10.0	0.0415

Overall Comparison of the Two Analytical Tracks:

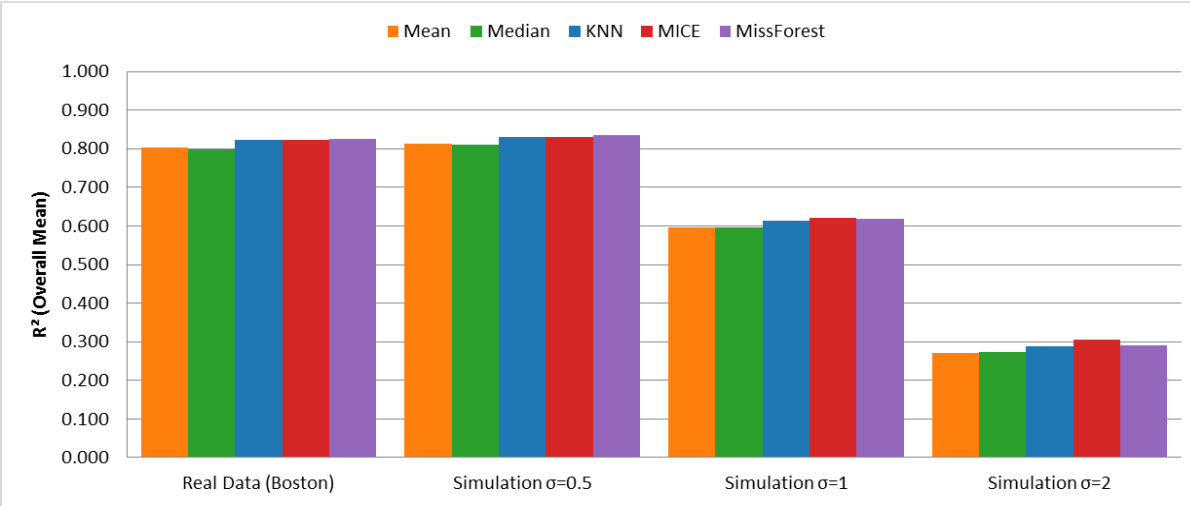


Figure (6): Comparison of the Coefficient of Determination (R^2) Between the Real Data and the Three Noise Levels in the Simulated Data

Based on the comparison, it can be seen that in general, MissForest and MICE were the best relative performances. The overall mean performance achieved by combining SVR with MissForest was the highest one (R^2=0.825) in Boston Housing data. MissForest also led under low noise simulation (R^2=0.836). As noise increased MICE performed the best under high noise (R^2=0.305), SVR's performance decreased significantly at (σ=2) and the effect was consistent for the various imputation methods.

Conclusion:

This paper considered the issue of missing at random for support vector regression. It has been found that SVR is sensitive to data quality. Its performance may decline when noise levels and proportions of missingness are high, even with the best-performing imputation methods. Accurate, efficient SVR prediction therefore depends on imputed data that preserve the original dataset's statistical structure adequately. Larger sample sizes give imputation methods more information, potentially improving missing-value estimates. Adding variables, on the other hand, increases the number of relationships those methods must estimate, which may make imputation more sensitive to higher proportions of missing values. MissForest and MICE, the more advanced approaches, performed better than the conventional methods in most comparisons. Their results indicate that using relationships among multiple variables helps estimate missing values more accurately than mean or median imputation. At the high-noise level, though, the methods performed much more similarly with MICE remaining the best, suggesting that substantial noise limits the extent to which imputation can improve model performance. Mean and median imputation were consistently the weakest across every simulation condition and in the real-data application, especially as the missingness rate increased, as both depend on a single measure of central tendency, which may fail to capture relationships between variables and may not preserve the data's correlation structure consistently. KNN, however, performed competitively with the advanced methods, slightly outperforming MICE on the real data at higher missingness rates. These results may give researchers some valuable guidelines for selecting their approaches in missing data management in Support Vector Regression (SVR) models.

References:

1. Batra, S., Khurana, R., Khan, M. Z., Boulila, W., Koubaa, A., & Srivastava, P. (2022). A pragmatic ensemble strategy for missing values imputation in health records. Entropy, 24(4), 533.
2. Ibrahim, N., & Wibowo, A. (2014). Support vector regression with missing data treatment based variables selection for water level prediction of Galas River in Kelantan Malaysia. WSEAS Transactions on Mathematics, 13, 69–78.
3. Meyer, C., Baogui, D., & Gouda, M. A. (2025). Applying machine learning to gauge the number of women in science, technology, and innovation policy (STIP): A model to accommodate missing data. Humanities and Social Sciences Communications, 12(1), 1245.
4. Shams, M. Y., Elshewey, A. M., El-Kenawy, E. S. M., Ibrahim, A., Talaat, F. M., & Tarek, Z. (2024). Water quality prediction using machine learning models based on grid search method. Multimedia Tools and Applications, 83, 35307–35334.
5. El Badisy, I., Graffeo, N., Khalis, M., & Giorgi, R. (2024). Multi-metric comparison of machine learning imputation methods with application to breast cancer survival. BMC Medical Research Methodology, 24, 191.
6. Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. Advances in Neural Information Processing Systems, 9, 155–161.
7. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199–222.
8. Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581–592.
9. Enders, C. K. (2022). Applied missing data analysis. Guilford Publications.
10. Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.
11. Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67.

12. Stekhoven, D. J., & Bühlmann, P. (2012). MissForest: Non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112–118.
13. Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. Journal of Environmental Economics and Management, 5(1), 81–102.

Appendix A: Raw simulation Results:

Note: values are averaged over the 500 replications and three sample sizes (N=50, 100, and 200) (RMSE, MAE, and R^2).

Table (A1): Raw simulation Results (RMSE) of the imputation methods by Noise Level, number of predictors, and missing rate

Noise (σ)	Predictors	Missing Rate	Mean	Median	KNN	MICE	MissForest
0.5	2	5%	0.541	0.542	0.537	0.535	0.536
		10%	0.547	0.547	0.539	0.537	0.539
		20%	0.573	0.572	0.549	0.554	0.544
		30%	0.614	0.614	0.569	0.588	0.565
	3	5%	0.568	0.567	0.559	0.554	0.556
		10%	0.584	0.585	0.560	0.558	0.557
		20%	0.657	0.657	0.601	0.608	0.587
		30%	0.714	0.713	0.634	0.655	0.615
	5	5%	0.624	0.626	0.604	0.602	0.600
		10%	0.687	0.689	0.642	0.644	0.628
		20%	0.800	0.802	0.707	0.711	0.668
		30%	0.947	0.951	0.790	0.817	0.731
1.0	2	5%	1.046	1.046	1.045	1.037	1.042
		10%	1.057	1.057	1.047	1.043	1.047
		20%	1.088	1.086	1.067	1.067	1.067
		30%	1.101	1.099	1.075	1.077	1.071
	3	5%	1.079	1.077	1.073	1.066	1.073
		10%	1.097	1.099	1.083	1.079	1.082
		20%	1.136	1.138	1.098	1.094	1.090
		30%	1.204	1.199	1.137	1.143	1.121
	5	5%	1.143	1.142	1.127	1.118	1.121

Noise (σ)	Predictors	Missing Rate	Mean	Median	KNN	MICE	MissForest
		10%	1.184	1.186	1.157	1.143	1.145
		20%	1.271	1.275	1.194	1.193	1.180
		30%	1.381	1.374	1.260	1.256	1.219
2.0	2	5%	2.085	2.083	2.080	2.070	2.082
		10%	2.093	2.090	2.084	2.070	2.085
		20%	2.095	2.103	2.084	2.074	2.088
		30%	2.136	2.138	2.115	2.102	2.115
	3	5%	2.138	2.134	2.131	2.114	2.132
		10%	2.125	2.124	2.111	2.091	2.104
		20%	2.165	2.157	2.142	2.115	2.130
		30%	2.206	2.199	2.160	2.140	2.153
	5	5%	2.193	2.192	2.179	2.163	2.184
		10%	2.204	2.201	2.179	2.152	2.170
		20%	2.282	2.285	2.225	2.199	2.215
		30%	2.364	2.368	2.286	2.252	2.251

Table (A2): Raw simulation results (MAE) of the imputation methods by Noise Level, number of predictors, and missing rate

Noise (σ)	Predictors	Missing Rate	Mean	Median	KNN	MICE	MissForest
0.5	2	5%	0.434	0.434	0.431	0.430	0.431
		10%	0.439	0.439	0.432	0.431	0.432
		20%	0.454	0.454	0.437	0.441	0.434
		30%	0.484	0.485	0.453	0.468	0.451
	3	5%	0.452	0.452	0.445	0.443	0.444
		10%	0.464	0.465	0.447	0.446	0.445
		20%	0.517	0.518	0.477	0.483	0.468
		30%	0.559	0.559	0.501	0.518	0.488
	5	5%	0.495	0.498	0.481	0.479	0.477
		10%	0.543	0.545	0.509	0.510	0.498
		20%	0.626	0.629	0.558	0.561	0.529
		30%	0.737	0.741	0.618	0.641	0.573
1.0	2	5%	0.842	0.842	0.841	0.835	0.839
		10%	0.850	0.851	0.843	0.840	0.843
		20%	0.875	0.873	0.858	0.859	0.857
		30%	0.884	0.882	0.864	0.866	0.860
	3	5%	0.867	0.867	0.863	0.858	0.862
		10%	0.882	0.882	0.871	0.868	0.870
		20%	0.912	0.914	0.883	0.882	0.877
		30%	0.963	0.961	0.913	0.918	0.901
	5	5%	0.919	0.918	0.905	0.899	0.901
		10%	0.948	0.950	0.927	0.917	0.918
		20%	1.017	1.018	0.954	0.953	0.942
		30%	1.102	1.097	1.010	1.005	0.977
2.0	2	5%	1.679	1.675	1.673	1.665	1.674
		10%	1.684	1.682	1.676	1.666	1.678
		20%	1.685	1.691	1.678	1.669	1.681
		30%	1.719	1.720	1.701	1.691	1.702
	3	5%	1.722	1.719	1.714	1.703	1.716

Noise (σ)	Predictors	Missing Rate	Mean	Median	KNN	MICE	MissForest
		10%	1.713	1.713	1.701	1.686	1.696
		20%	1.744	1.737	1.724	1.704	1.715
		30%	1.776	1.771	1.737	1.723	1.732
	5	5%	1.763	1.763	1.752	1.740	1.756
		10%	1.772	1.771	1.755	1.731	1.745
		20%	1.832	1.837	1.789	1.767	1.783
		30%	1.896	1.900	1.835	1.810	1.808

Table (A3): Raw simulation results (R^2) of the imputation methods by Noise Level, number of predictors, and missing rate

Noise (σ)	Predictors	Missing Rate	Mean	Median	KNN	MICE	MissForest
0.5	2	5%	0.791	0.792	0.795	0.796	0.795
		10%	0.780	0.780	0.787	0.788	0.786
		20%	0.755	0.753	0.770	0.769	0.773
		30%	0.752	0.751	0.778	0.770	0.778
	3	5%	0.849	0.849	0.853	0.856	0.854
		10%	0.843	0.844	0.855	0.857	0.856
		20%	0.810	0.807	0.829	0.832	0.839
		30%	0.773	0.773	0.809	0.803	0.818
	5	5%	0.892	0.891	0.900	0.901	0.901
		10%	0.871	0.870	0.887	0.887	0.891
		20%	0.836	0.832	0.864	0.865	0.880
		30%	0.785	0.782	0.837	0.826	0.856
1.0	2	5%	0.544	0.544	0.546	0.553	0.548
		10%	0.534	0.535	0.542	0.547	0.541
		20%	0.506	0.508	0.522	0.526	0.522
		30%	0.479	0.481	0.498	0.506	0.502
	3	5%	0.620	0.621	0.624	0.631	0.622
		10%	0.612	0.611	0.620	0.626	0.620
		20%	0.600	0.598	0.621	0.629	0.624
		30%	0.566	0.571	0.605	0.609	0.616
	5	5%	0.719	0.720	0.727	0.733	0.730
		10%	0.705	0.703	0.718	0.728	0.723
		20%	0.653	0.651	0.689	0.693	0.693
		30%	0.598	0.604	0.654	0.663	0.673
2.0	2	5%	0.220	0.222	0.225	0.234	0.222
		10%	0.193	0.195	0.202	0.214	0.200
		20%	0.206	0.202	0.215	0.226	0.211
		30%	0.183	0.185	0.202	0.216	0.200
	3	5%	0.274	0.279	0.281	0.294	0.279
		10%	0.269	0.267	0.278	0.295	0.283
		20%	0.261	0.273	0.278	0.303	0.288
		30%	0.241	0.247	0.269	0.291	0.276
	5	5%	0.383	0.384	0.392	0.406	0.388
		10%	0.379	0.380	0.395	0.412	0.399
		20%	0.351	0.348	0.377	0.400	0.385
		30%	0.301	0.294	0.337	0.364	0.356