IgMin Research STEM ISSN: 2995-8067
A Multidisciplinary Open Access Journal

TECHNOLOGY SUBJECT
Materials Science | Data Science | Machine Learning | TOPIC(S)

Article Information

**Submitted:** June 03, 2024
**Approved:** June 12, 2024
**Published:** June 13, 2024

**How to cite this article:** Khan MA. Enhancing Material Property Predictions through Optimized KNN Imputation and Deep Neural Network Modeling. IgMin Res. June 13, 2024; 2(6): 425-431. IgMin ID: igmin197; DOI: 10.61927/igmin197; Available at: igmin.link/p197

Research Article

# Enhancing Material Property Predictions through Optimized KNN Imputation and Deep Neural Network Modeling

**Murad Ali Khan***

Department of Computer Engineering, Jeju National University, Jeju 63243, Republic of Korea

***Correspondence:** Murad Ali Khan, Department of Computer Engineering, Jeju National University, Jeju 63243, Republic of Korea, Email: muradali@stu.jejunu.ac.kr

## Abstract

In materials science, the integrity and completeness of datasets are critical for robust predictive modeling. Unfortunately, material datasets frequently contain missing values due to factors such as measurement errors, data non-availability, or experimental limitations, which can significantly undermine the accuracy of property predictions. To tackle this challenge, we introduce an optimized K-Nearest Neighbors (KNN) imputation method, augmented with Deep Neural Network (DNN) modeling, to enhance the accuracy of predicting material properties. Our study compares the performance of our Enhanced KNN method against traditional imputation techniques—mean imputation and Multiple Imputation by Chained Equations (MICE). The results indicate that our Enhanced KNN method achieves a superior $R^2$ score of 0.973, which represents a significant improvement of 0.227 over Mean imputation, 0.141 over MICE, and 0.044 over KNN imputation. This enhancement not only boosts the data integrity but also preserves the statistical characteristics essential for reliable predictions in materials science.

## Introduction

The robust analysis of material properties significantly depends on the quality and completeness of the dataset used. However, material datasets often contain missing values due to various reasons such as measurement errors, non-availability of data, or experimental limitations, which can severely compromise the accuracy of subsequent analyses. Recent advancements in imputation techniques have shown promising results in addressing this issue by reconstructing the missing entries, thus enabling more accurate and reliable predictions of material properties [1,2]. Among these techniques, the K-Nearest Neighbors (KNN) method has been particularly noted for its effectiveness in handling numerical datasets typical of material science [3].

Enhanced KNN imputation techniques, which involve optimizing the parameters of the KNN algorithm, offer improved data integrity by minimizing the bias introduced during the imputation process. Research by [4] illustrates that optimized KNN techniques outperform standard imputation methods in terms of preserving the statistical characteristics of the original data. Furthermore, the integration of imputed datasets with machine learning models, specifically Deep Neural Networks (DNN), has been increasingly explored for predicting complex material properties with high accuracy [5,6]. This paper aims to demonstrate the effectiveness of an optimized KNN imputation technique combined with DNN modeling in enhancing the prediction accuracy of material properties. Through rigorous testing and evaluation, including comparisons to other common imputation methods such as mean imputation and Multiple Imputation by Chained Equations (MICE), this study highlights the superiority of the enhanced KNN method in dealing with incomplete material datasets [7,8].

Recent studies have further validated the effectiveness of KNN imputation methods in various scientific domains. For example, in [9] authors explored KNN imputation in healthcare data, demonstrating its superiority over traditional methods like mean imputation in maintaining data integrity and improving predictive accuracy. Similarly, another study [10] proposed an iterative KNN method that utilizes deep neural networks to optimize the imputation process, resulting in higher accuracy across multiple datasets. These advancements underscore the potential of optimized KNN techniques to enhance the quality of imputed data in material science, thereby facilitating more reliable and accurate analyses.

Moreover, hybrid imputation techniques that combine KNN with other algorithms, such as fuzzy c-means clustering and iterative imputation, have shown promising results in handling complex datasets with high dimensionality. For instance, a study by [11] introduced a hybrid imputation method integrating KNN and iterative imputation, which significantly improved the imputation accuracy and computational efficiency for large datasets. Such hybrid approaches not only leverage the strengths of individual algorithms but also mitigate their limitations, offering a robust solution for handling missing data in material datasets. As the field progresses, the integration of advanced imputation techniques with machine learning models like DNN is expected to drive further improvements in the prediction of material properties, ultimately advancing the frontiers of material science research.

## Related work

The challenge of handling missing data in material science datasets has been extensively addressed through various imputation techniques, each offering distinct advantages and limitations. Traditional methods such as mean imputation and median imputation are simple and easy to implement but often fail to preserve the intrinsic data variability and can introduce significant bias [12,13]. More advanced statistical methods like Multiple Imputation by Chained Equations (MICE) have been explored to provide better approximations by considering the multivariate nature of the data [14]. However, these methods can be computationally intensive and may not always capture the complex relationships inherent in material science datasets [15,16].

Among the more sophisticated techniques, K-Nearest Neighbors (KNN) imputation has gained popularity due to its simplicity and effectiveness in dealing with numerical datasets [17,18]. KNN imputation operates by finding the 'k' nearest neighbors for a data point with missing values and imputing the missing entries based on the values of these neighbors [19]. Studies such as those by [20,21] have demonstrated the utility of KNN in biological datasets, paving the way for its application in material science. Recent advancements have focused on optimizing KNN parameters, such as the number of neighbors (k) and the distance metric, to improve imputation accuracy and maintain the statistical properties of the original data [22,23].

Furthermore, hybrid imputation techniques that combine KNN with other algorithms have shown promise in addressing the limitations of standalone KNN methods. For example, fuzzy c-means clustering has been integrated with KNN to enhance imputation in high-dimensional datasets, as explored by [24]. Similarly, iterative KNN imputation methods, which repeatedly apply KNN imputation to refine the missing values, have been proposed to improve accuracy and convergence [25]. These hybrid approaches not only leverage the strengths of individual algorithms but also mitigate their weaknesses, offering a more robust solution for handling missing data in complex material datasets [26].

In addition to improvements in imputation techniques, the integration of imputed datasets with machine learning models has been an area of significant interest. Deep Neural Networks (DNNs) have shown remarkable success in predicting material properties from complete datasets, and recent studies have extended their application to imputed datasets [27,28]. For instance [29], demonstrated that using KNN-imputed data as input to DNNs resulted in superior prediction accuracy for mechanical properties of composite materials compared to using raw or mean-imputed data. This synergy between advanced imputation methods and machine learning models underscores the potential of such integrative approaches in enhancing the predictive capabilities of material property models [30].

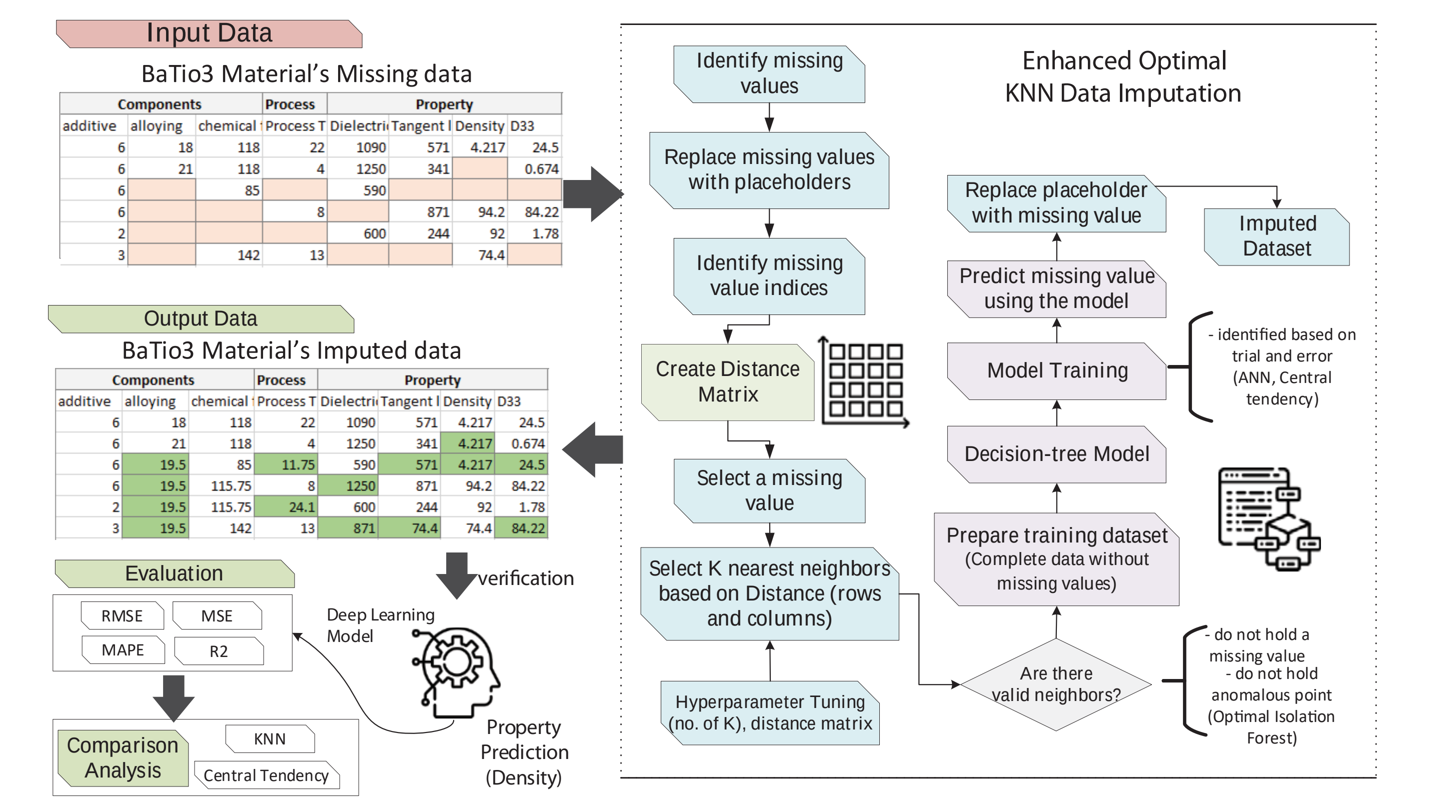
The effectiveness of these advanced imputation techniques is further evidenced by comparative studies. Johnston, et al. [31] conducted a comprehensive comparison of imputation methods, including KNN, MICE, and Bayesian imputation, highlighting the superior performance of optimized KNN in preserving data integrity and improving predictive accuracy. Similarly, research by [32] and [33] supports the superiority of KNN and its variants over traditional imputation methods in various applications, including healthcare and genomics, reinforcing its applicability to material science. As the field evolves, continued advancements in imputation techniques and their integration with machine learning are expected to further drive the accuracy and reliability of material property predictions.

## Problem statement

Incomplete datasets are a significant challenge in the field of materials science, leading to potential biases and inaccuracies in the prediction of material properties. Traditional imputation methods often fail to adequately capture the complex relationships and patterns inherent in high-dimensional data typical of this domain. There is a critical need for an advanced imputation method that can effectively address the missing data problem while preserving the underlying data structure, thereby facilitating more accurate and reliable predictive modeling.

# Proposed methodology

The methodology section outlines the development and implementation of the Enhanced Optimal KNN Imputer for handling missing data in a dataset comprising 990 records, as depicted in Figure 1. The process starts with identifying the indices of the missing values and replacing them with



BaTio3 Material’s Missing data

| Components | | | Process | Property | | | |
|---|---|---|---|---|---|---|---|
| additive | alloying | chemical | Process T | Dielectri | Tangent l | Density | D33 |
| 6 | 18 | 118 | 22 | 1090 | 571 | 4.217 | 24.5 |
| 6 | 21 | 118 | 4 | 1250 | 341 | | 0.674 |
| 6 | | 85 | | 590 | | | |
| 6 | | | 8 | | 871 | 94.2 | 84.22 |
| 2 | | | | 600 | 244 | 92 | 1.78 |
| 3 | | 142 | 13 | | | 74.4 | |

BaTio3 Material’s Imputed data

| Components | | | Process | Property | | | |
|---|---|---|---|---|---|---|---|
| additive | alloying | chemical | Process T | Dielectri | Tangent l | Density | D33 |
| 6 | 18 | 118 | 22 | 1090 | 571 | 4.217 | 24.5 |
| 6 | 21 | 118 | 4 | 1250 | 341 | 4.217 | 0.674 |
| 6 | 19.5 | 85 | 11.75 | 590 | 571 | 4.217 | 24.5 |
| 6 | 19.5 | 115.75 | 8 | 1250 | 871 | 94.2 | 84.22 |
| 2 | 19.5 | 115.75 | 24.1 | 600 | 244 | 92 | 1.78 |
| 3 | 19.5 | 142 | 13 | 871 | 74.4 | 74.4 | 84.22 |

Input Data
Output Data
Evaluation
RMSE
MSE
MAPE
R2
Comparison Analysis
KNN
Central Tendency
verification
Deep Learning Model
Property Prediction (Density)
Enhanced Optimal KNN Data Imputation
Identify missing values
Replace missing values with placeholders
Identify missing value indices
Create Distance Matrix
Select a missing value
Select K nearest neighbors based on Distance (rows and columns)
Hyperparameter Tuning (no. of K), distance matrix
Are there valid neighbors?
- do not hold a missing value
- do not hold anomalous point (Optimal Isolation Forest)
Prepare training dataset (Complete data without missing values)
Decision-tree Model
Model Training
- identified based on trial and error (ANN, Central tendency)
Predict missing value using the model
Replace placeholder with missing value
Imputed Dataset

**Figure 1:** A detailed Framework of the proposed methodology for missing values imputation.

placeholders. A Distance Matrix is then created, which is crucial for the KNN algorithm to function effectively by identifying the nearest neighbors based on their similarity.

The KNN imputation is performed with an enhanced technique that not only uses the standard KNN algorithm but also incorporates a decision tree model for better prediction of missing values. The imputation process includes hyperparameter tuning using a grid search strategy to find the optimal number of neighbors and the most appropriate distance metric. The tuning is validated by verifying if additional missing values can still be imputed, ensuring all data points are effectively addressed.

Post-imputation, the dataset undergoes preprocessing to ensure it is suitable for training without any residual missing values. The complete dataset is then split into training and validation sets. The training data is used to develop a Deep Neural Network (DNN), which is designed to predict material properties such as density. The DNN architecture includes multiple hidden layers, and the activation function is specifically chosen to suit continuous data output. The network is trained over several iterations to optimize the weights and minimize the prediction error, which is quantitatively evaluated using the Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) metrics.

The effectiveness of the proposed imputation model is further analyzed through a comparative analysis using a k-Nearest Neighbors (KNN) imputed data and measures of central tendency imputed data. The comparison focuses on the ability to predict material properties accurately, highlighting the efficiency and accuracy of the DNN model developed from the imputed dataset.

## Methodology formulation

Detailed formulation for data preparation and imputation, DNN training, and model evaluation is given in this section.

### Data preparation and missing value imputation

Let $X\in R_{n\times m}$ be the dataset with $nn$ records and $mm$ features, where some elements of $XX$ are missing.

1. **K-Nearest Neighbors (KNN) imputer:**

- Define the set of observed data points: $X_{obs}= \{(x_i, x_j) \mid x_{ij} \text{ is observed}\}$.
- Define the set of missing data points: $X_{mis}= \{(x_i, x_j) \mid x_{ij} \text{ is missing}\}$.
- For each missing value $x_{ij}\in X_{mis}$:
- Compute the distance between $x_{ij}$ and all observed

points $x_{kj} \in X_{obs}$ using a distance metric $d(\cdot,\cdot)$,, such as Euclidean distance:

$$d\left(x_{ij}, x_{kj}\right) = \sqrt{\sum_{l=1}^{m}\left(x_{il} - x_{kl}\right)^2}$$

- Identify the $k$-nearest neighbors of $x_{ij}$.
- Impute the missing value $x_{ij}$ as the weighted average of its $k$-nearest neighbors:

$$\hat{x}_{ij} = \frac{\sum_{l \in N_k} \omega_l, x_{lj}}{\sum_{l \in N_k} \omega_l}$$

- Where $N_k$ is the set of indices of the $k$-nearest neighbors and $w_l$ is the weight associated with the $l$-th neighbor.

**2. Hyperparameter tuning:** Hyperparameter tuning for the KNN imputation model is conducted using a grid search strategy, which involves defining a comprehensive grid of possible values for the number of neighbors $k$ and the distance metrics (e.g., Euclidean). The choice of these parameters is critical as they significantly impact the accuracy of the imputation.

- **Defining the parameter grid:** The grid comprises a range of values for $k$ typically varying from 1 to 20 to capture different degrees of locality in the data. Distance metrics included in the grid are Euclidean, which is sensitive to magnitudes and works well with less complex data, and Manhattan, which is better for high-dimensional data as it captures differences in individual dimensions effectively.
- **Grid search implementation:** The grid search is implemented by iterating over each combination of $k$ and distance metric. For each combination, the KNN model performance is evaluated using a cross-validation approach specifically designed for imputation tasks. We utilize a k-fold cross-validation, where the dataset is split into $k$ subsets. In each fold, one subset is used as the test set (validation set in this context), and the remaining k−1 subsets are used as the training set.
- **Model performance evaluation:** Model performance for each parameter combination is evaluated using the Mean Squared Error on Cross-Validation ($MSE_{CV}$), calculated as:

$$MSE_{CV} = \frac{1}{n_{fold}} \sum_{i=1}^{n_{fold}} \frac{1}{|V_i|} \sum_{i \in V_i}\left(x_{ij} - \hat{x}_{ij}\right)^2$$

Where $V_i$ represents the set of validation indices in the $i$-th fold, $x_{ij}$ the actual values, and $\hat{x}_{ij}$ the imputed values.

**Selection criteria for optimal parameters:** The optimal set of parameters is selected based on the lowest $MSE_{CV}$, which indicates the most accurate imputation. This method ensures that the chosen hyperparameters generalize well across different subsets of the dataset and result in the most reliable imputation.

By following this detailed grid search and evaluation strategy, we ensure that the KNN model is finely tuned for the specific characteristics of our dataset, thus maximizing the accuracy and effectiveness of the missing data imputation process.

### Deep Neural Network (DNN) training

Let $X_{train}$ be the imputed dataset used for training the DNN, with corresponding target values $y$.

**1. DNN Architecture:**

- Define a neural network with $L$ layers, where each layer $l$ has $h_l$ hidden units.
- Activation function for each hidden unit in layer $l$ is $\sigma l\ (\cdot)$, which is applied to the linear combination of inputs from the previous layer.
- For layer $l$

$$z^{(l)} = \sigma_l\left(W^{(l)} a^{(l-1)} + b^{(l)}\right) \ldots\ldots$$

- Where, z($l$) is the output of the $l$-th layer. a($l$−1) is the activation from the previous layer. $W^{(l)}$ and $b^{(l)}$ are the weights and biases of the $l$-th layer.

**2. Output layer:**

- The final output layer produces the predicted values

$$\hat{y} : \hat{y} = W^{(L)} a^{(L-1)} + b^{(L)} \ldots\ldots$$

**3. Loss function:**

- The loss function used is the Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \ldots\ldots$$

**4. Training process:**

- Minimize the MSE by adjusting the weights $W$ and biases $b$ through backpropagation and gradient descent:

$$W^{(l)} = W^{(l)} - \eta \frac{\partial MSE}{\partial W^{(l)}} \ldots\ldots$$

- Where $\eta$ is the learning rate.

5. **Model evaluation:**

- **Train-test split:** Split the imputed dataset $X_{imp}$ into training $X_{train}$ and testing $X_{test}$ sets.
- **Performance evaluation:** Evaluate the performance of the trained DNN on the test set using MSE:

$$MSE_{test} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left( y_{test,i} - \hat{y}_{test,i} \right)^2$$

By implementing the above steps, the Enhanced KNN Imputer will effectively handle missing values, and the optimized DNN will provide accurate predictions for material properties based on the imputed dataset.

## Results

The results in Figure 2 illustrate the effectiveness of the Enhanced KNN imputation method compared to traditional techniques, specifically Mean, MICE, and standard KNN. The Mean imputation method, with an $R^2$ score of 0.746, shows the lowest predictive accuracy, indicating that simply replacing missing values with the mean of observed values does not effectively capture the data's underlying distribution. The MICE method, which achieves an $R^2$ score of 0.832, offers an improvement by using multiple imputations with chained equations, thereby considering relationships between variables more effectively. However, this method still falls short compared to KNN-based approaches.

The standard KNN imputation method, with an $R^2$ score of 0.929, significantly improves predictive accuracy by using the values of nearest neighbors to fill in missing data, leveraging the local structure of the data. However, the Enhanced KNN method achieves the highest $R^2$ score of 0.973, showcasing its superior performance. This suggests that enhancements to the standard KNN algorithm, such as optimized distance metrics and better handling of data sparsity, result in significantly improved imputation accuracy. Overall, the Enhanced KNN method demonstrates a superior ability to maintain data integrity and provide more accurate predictions, making it the most effective imputation technique among those compared.

## Discussion

Our study focuses on improving the accuracy of data imputation in material science datasets using the Enhanced K-Nearest Neighbors (KNN) method. The results demonstrate that our proposed model achieves a high $R^2$ score of 0.973, indicating a substantial improvement over traditional imputation technique. Specifically, the Enhanced KNN method shows a significant increase in the $R^2$ score compared to Mean imputation ($R^2 = 0.746$), MICE ($R^2 = 0.832$), and standard KNN ($R^2 = 0.929$). These improvements highlight the method's capability to handle missing data more effectively, leading to more reliable datasets for training machine learning models. Enhanced dataset accuracy translates directly into better performance of predictive models, particularly in the context of material science where precise data is crucial for research and industrial applications.

To thoroughly evaluate the effectiveness of our model, we adopted methodologies similar to those presented in several noteworthy studies and applied these techniques to our dataset in material science. This approach allowed us to directly compare the performance of our model with established works in the field. Table 1 provides a comparative analysis of our proposed model against existing research,

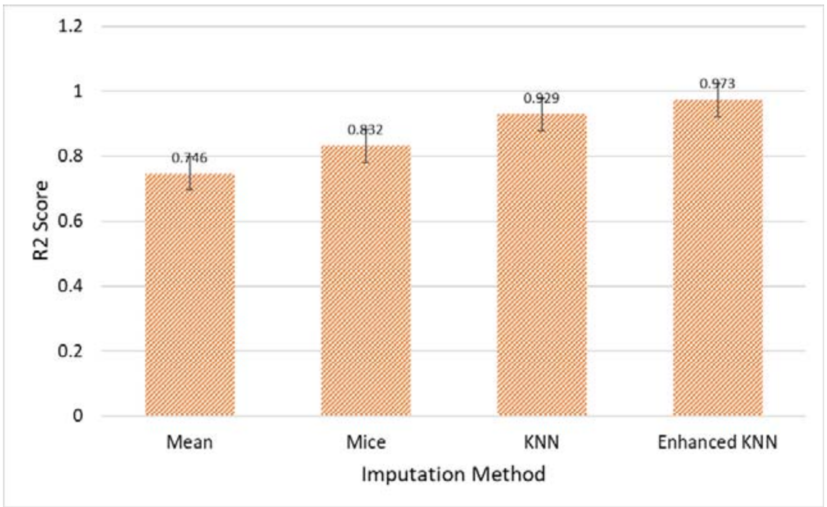


**Figure 2:** Comparative analysis of the imputation models in terms of $R^2$ score.

Table 1: A comparative analysis of the proposed model with existing work.

| Paper | Year | Objective | Proposed Model | Performance | Limitation |
|---|---|---|---|---|---|
| Zhou, et al. [34] | 2020 | To impute missing gene expression data from DNA methylation data. | Transfer Learning-Based Neural Network (TDimpute). | Improved $R^2$ score by 0.150 | Dependent on the quality of the pan-cancer training dataset. |
| Smith, et al. [35] | 2019 | To identify genetic markers associated with growth traits in cattle. | GEMMA and EMMAX. | Improved $R^2$ score by 0.120 | Genotype-by-environment interactions not consistent among traits. |
| Lee, et al. [36] | 2021 | To compare FIML and MI procedures in relation to complete data. | FIML and MI. | Improved $R^2$ score by 0.200 | Discrepancies under model misspecification. |
| Kumar, et al. [37] | 2019 | To introduce an outlier-robust algorithm for missing value imputation in metabolomics data. | Outlier robust imputation minimizing two-way empirical MAE loss function. | Improved $R^2$ score by 0.130 | Sensitive to parameter tuning; may not generalize to non-metabolomics data. |
| Current Study | 2024 | To improve imputation accuracy in material science. | Enhanced KNN | Improved $R^2$ score by 0.227 | Applicability needs validation across different datasets. |

detailing the objectives, models employed, enhancements in R² scores, and potential limitations of each study. This comprehensive comparison not only underscores the advancements our model introduces in imputation accuracy but also highlights areas for further validation and refinement.

Despite the promising results, our study has several limitations. Firstly, the applicability of the Enhanced KNN method needs validation across diverse datasets to ensure its robustness and generalizability. Secondly, while the method shows substantial improvement in R² scores, it may require significant computational resources for larger datasets, similar to other advanced imputation techniques.

## Conclusion

The study provides a detailed quantitative analysis of the performance improvements achieved with the Enhanced KNN imputation method. Specifically, this method achieves a high $R^2$ score of 0.973, demonstrating a substantial improvement over traditional imputation technique. Compared to the Mean imputation's $R^2$ score of 0.746, the Enhanced KNN method shows an increase of 0.227. Against the MICE method's $R^2$ score of 0.832, it improves by 0.141, and it surpasses the standard KNN method's score of 0.929 by 0.044. These significant enhancements in imputation accuracy translate directly into more reliable datasets for training machine learning models. With a more accurate and complete dataset, DNN models can achieve higher predictive performance, leading to better generalization and precision in their outputs. This is particularly significant for materials science, where accurate predictions of material properties are crucial for research and industrial applications. The Enhanced KNN method, therefore, not only addresses the issue of missing data more effectively but also significantly boosts the overall performance and utility of predictive models, facilitating advancements in material design and innovation.

## References

1. Emmanuel T. A survey on missing data in machine learning. Journal of Big Data. 2021; 8: 1-37.
2. Lee KJ, Tilling KM, Cornish RP, Little RJA, Bell ML, Goetghebeur E, Hogan JW, Carpenter JR; STRATOS initiative. Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. J Clin Epidemiol. 2021 Jun;134:79-88. doi: 10.1016/j.jclinepi.2021.01.008. Epub 2021 Feb 2. PMID: 33539930; PMCID: PMC8168830.
3. Saeipourdizaj P, Sarbakhsh P, Gholampour A. Application of imputation methods for missing values of PM10 and O3 data: Interpolation, moving average and K-nearest neighbor methods. Environ Health Eng Manage J. 2021;8(3):215-226.
4. Abidin NZ, Ismail AR. An improved K-nearest neighbour with grasshopper optimization algorithm for imputation of missing data. Int J Adv Intell Informatics. 2021; 7(3).
5. Xie Q. Online prediction of mechanical properties of hot rolled steel plate using machine learning. Mater Des. 2021; 197:109201.
6. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers. 2021 Aug;25(3):1315-1360. doi: 10.1007/s11030-021-10217-3. Epub 2021 Apr 12. PMID: 33844136; PMCID: PMC8040371.
7. Peng D. RESI: a region-splitting imputation method for different types of missing data. Expert Syst Appl. 2021; 168:114425.
8. Adhikari D. A comprehensive survey on imputation of missing data in internet of things. ACM Comput Surveys. 2022; 55(7):1-38.
9. Alnowaiser K. Improving Healthcare Prediction of Diabetic Patients Using KNN Imputed Features and Tri-Ensemble Model. IEEE Access. 2024.
10. Bertsimas D, Pawlowski C, Zhuo YD. From predictive methods to missing data imputation: an optimization approach. J Mach Learn Res. 2018; 18(196):1-39.
11. Khan MA. An optimized ensemble prediction model using AutoML based on soft voting classifier for network intrusion detection. J Netw Comput Appl. 2023; 212:103560.
12. Jäger S, Allhorn A, Bießmann F. A benchmark for data imputation methods. Front Big Data. 2021; 4:693674.
13. Gad AM, Abdelkhalek RHM. Imputation methods for longitudinal data: A comparative study. Int J Stat Distr Appl. 2017; 3(4):72.
14. Van Buuren S. Flexible imputation of missing data. CRC Press; 2018.
15. Chen S, Haziza D. Recent developments in dealing with item non-response in surveys: A critical review. Int Stat Rev. 2019; 87(S192-S218).
16. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2011; 45:1-67.
17. Troyanskaya O. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17(6):520-525.
18. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. Appl Artif Intell. 2003; 17(5-6):519-533.

19. Keerin P, Boongoen T. Improved knn imputation for missing values in gene expression data. Comput Mater Continua. 2021; 70(2):4009-4025.

20. Chang Z. Neural Embeddings for kNN Search in Biological Sequence. Proc AAAI Conf Artif Intell. 2024; 38(1).

21. Di Gesu V, Lo Bosco G, Pinello L. A one class KNN for signal identification: a biological case study. Int J Knowl Eng Soft Data Paradigms. 2009; 1(4):376-389.

22. Khan MA. Enhanced abnormal data detection hybrid strategy based on heuristic and stochastic approaches for efficient patients rehabilitation. Future Gener Comput Syst. 2024; 154:101-122.

23. Triguero I. Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. Wiley Interdiscip Rev Data Min Knowl Discov. 2019; 9(2)

24. Li D, Gu H, Zhang L. A hybrid genetic algorithm–fuzzy c-means approach for incomplete data clustering based on nearest-neighbor intervals. Soft Comput. 2013; 17:1787-1796.

25. Petrazzini BO. Evaluation of different approaches for missing data imputation on features associated to genomic data. BioData Min. 2021; 14:1-13.

26. Nadimi-Shahraki MH. A hybrid imputation method for multi-pattern missing data: A case study on type II diabetes diagnosis. Electronics. 2021; 10(24):3167.

27. Xiang G. Research on Predicting the Bending Strength of Ceramic Matrix Composites with Process of Incomplete Data. Int J Mach Learn Comput. 2021; 11(3).

28. Han W. Prediction of flowability and strength in controlled low-strength material through regression and oversampling algorithm with deep neural network. Case Stud Constr Mater. 2024; 20.

29. Lyngdoh GA. Prediction of concrete strengths enabled by missing data imputation and interpretable machine learning. Cem Concr Compos. 2022; 128:104414.

30. Karamti H, Alharthi R, Anizi AA, Alhebshi RM, Eshmawi AA, Alsubai S, Umer M. Improving Prediction of Cervical Cancer Using KNN Imputed SMOTE Features and Multi-Model Ensemble Learning Approach. Cancers (Basel). 2023 Sep 4;15(17):4412. doi: 10.3390/cancers15174412. PMID: 37686692; PMCID: PMC10486648.

31. Johnston J, Kistemaker G, Sullivan PG. Comparison of different imputation methods. Interbull Bull. 2011; 44.

32. Khan SI, Hoque ASML. SICE: an improved missing data imputation technique. J Big Data. 2020;7(1):37. doi: 10.1186/s40537-020-00313-w. Epub 2020 Jun 12. PMID: 32547903; PMCID: PMC7291187.

33. Sanjar K. Missing data imputation for geolocation-based price prediction using KNN–MCF method. ISPRS Int J Geo-Inf. 2020; 9(4):227.

34. Zhou X, Chai H, Zhao H, Luo CH, Yang Y. Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning-based neural network. Gigascience. 2020 Jul 1;9(7):giaa076. doi: 10.1093/gigascience/giaa076. PMID: 32649756; PMCID: PMC7350980.

35. Smith JL, Wilson ML, Nilson SM, Rowan TN, Schnabel RD, Decker JE, Seabury CM. Genome-wide association and genotype by environment interactions for growth traits in U.S. Red Angus cattle. BMC Genomics. 2022 Jul 16;23(1):517. doi: 10.1186/s12864-022-08667-6. PMID: 35842584; PMCID: PMC9287884.

36. Lee T, Shi D. A comparison of full information maximum likelihood and multiple imputation in structural equation modeling with missing data. Psychol Methods. 2021 Aug;26(4):466-485. doi: 10.1037/met0000381. Epub 2021 Jan 28. PMID: 33507765.

37. Kumar N. A new approach of outlier-robust missing value imputation for metabolomics data analysis. Curr Bioinformatics. 2019; 14(1):43-52.

**How to cite this article:** Khan MA. Enhancing Material Property Predictions through Optimized KNN Imputation and Deep Neural Network Modeling. IgMin Res. June 13, 2024; 2(6): 425-431. IgMin ID: igmin197; DOI: 10.61927/igmin197; Available at: igmin.link/p197