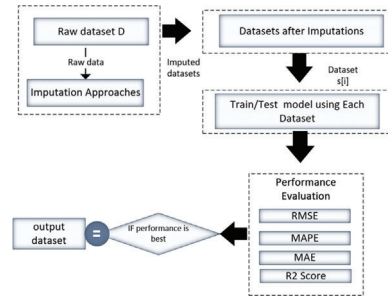


Imran,* N. Iqbal,
D.-H. Kim* 2200517

Intelligent Material Data Preparation Mechanism Based on Ensemble Learning for AI-Based Ceramic Material Analysis



This study presents intelligent material data preparation mechanism for the AI-based ceramic material analysis application. For the performance evaluation of the proposed methodology, machine learning models are trained before and after the imputation techniques applied to the ceramic data. Performance analysis shows that the ensemble model outperformed statistical imputation approaches.

Intelligent Material Data Preparation Mechanism Based on Ensemble Learning for AI-Based Ceramic Material Analysis

Imran,* Naeem Iqbal, and Do-Hyeun Kim*

Research interest in ceramic materials increased due to their extensive environmental, biomedical, and electronic applications. Increased demand for ceramics with specialized experimental conditions and limited resources has resulted in a higher cost for scientific practices and applications. Enormous material data is accumulated from traditional and high-tech experimentation, but the manual recording process has shown inconsistencies in the analysis results. Recently, application based on artificial intelligence (AI) and machine learning has been able to address the issues of traditional scientific experiments in material science. However, no machine learning mechanisms are proposed for sophisticated data preparation and AI-based discovery application of ceramics. This paper proposed an intelligent material data preparation mechanism based on ensemble learning for AI-assisted material screening and discovery. The current method can potentially resolve the problems of missing and inconsistent material data. As a case study, a material data preparation platform for ceramic material data pre-processing is developed. For performance evaluation of the proposed mechanism, machine learning regression models are trained before and after the imputation techniques applied to the data. Performance analysis shows that the ensemble model of deep learning network (DNN) and automated machine learning (autoML) performed better as compared to previously reported imputation approaches.

successful applications for high efficiency and performance, the need for AI in the simulation and exploration of novel ceramic materials is imperative. Furthermore, materials design based on AI analysis is foreseen to lead to innovative materials and reduce the development cost in terms of time and resources.^[5,6] However, the scientific community identified many limitations of advanced materials discovery and application. For instance, several issues related to computational simulation and the “materials” structures require high-performance index properties. Therefore, advanced materials research based on the convergence between AI techniques and experimental mechanisms is needed to produce the basic comprehension of the input parameters conditions and performance index properties.^[7]

It is prevalent in today's data science applications to encounter incomplete, missing, and insufficient data. Imputation is the solution used to address the missing or unknown data. Some previously used imputation methods fill missing values with the data feature's mean, max, median, mode, or minimum values. These techniques ignore

crucial contextual information. New intelligent techniques are needed to address the “imputation's goals to avoid bias in the imputed data. Thus, more sophisticated imputation techniques based on statistical inference and learning methods are imperative. These techniques will assist in determining missing values using the fitting function or mathematical model called interpolation. The datasets with missing data are problematic for many machine learning models. Data science experts resolved this using identifying and replacing missing values before model training for prediction and classification tasks.^[8–10] This mechanism is known as data imputation. A common approach for data imputation is calculating a statistical value for each data feature, such as a mean, and replacing the missing values for that column with the calculated value. Conversely, machine learning and data mining procedures are the most recently used approaches for data imputation. The main advantage of using data mining and machine learning for data imputation is extracting valuable knowledge from the complete data to develop prediction models for missing values prediction.^[11–13]

In literature, many algorithms such as K-nearest neighbor (KNN) and Bayesian network were used for data imputation.^[14–16] Hybrid mechanisms algorithms of decision

1. Introduction

AI applications are bringing drastic improvements in every field of life, especially in the technological fields.^[1–4] Due to its

Imran
Department of Biomedical Engineering
Gachon University
Incheon 21936, Republic of Korea
E-mail: imranj@gachon.ac.kr

N. Iqbal, D.-H. Kim
Department of Computer Engineering
Jeju National University
Jeju 63243, Republic of Korea
E-mail: kimdh@jeju.ac.kr

D.-H. Kim
Advanced Technology Research Institute
Jeju National University
Jeju 63243, Republic of Korea

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/adts.202200517>

DOI: 10.1002/adts.202200517

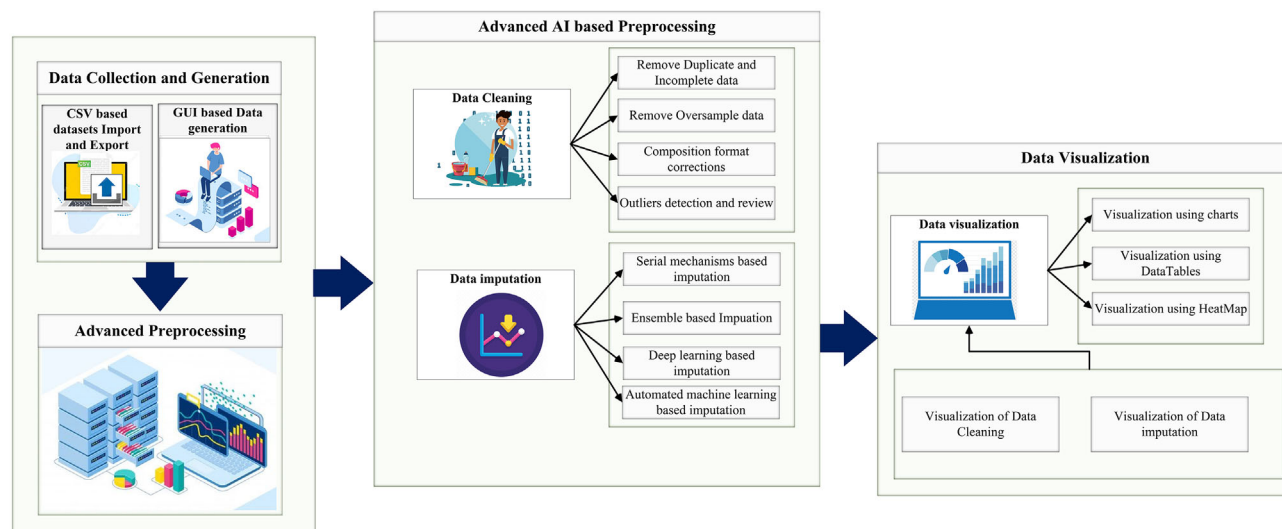


Figure 1. Conceptual design of the proposed intelligent data preparation mechanism.

tree algorithms and the maximum expectation algorithms are used for an improved version of data imputation algorithms.^[17,18] Study results show that hybrid approaches improve the classification accuracy of missing values filling. In literature, a data clustering mechanism was used for data imputation of categorical features.^[19,20] KNN-based imputation approach explores the entire dataset to obtain a data record closest to an assigned data record. It can identify data records with missing data, use the complete data records as the training dataset, and hunt for data records near the missing record. In the final step, data imputation of the missing data record is performed based on the nearest data record. The study results show that KNN has outstanding classification accuracy. Nevertheless, the KNN-based models have many shortcomings for the dataset with a large percentage of missing data, and missing values filling accuracy will significantly decrease.^[21,22]

Figure 1 presents the conceptual design of the proposed methodology. Advanced pre-processing is based on machine learning-based pre-processing such as data imputation and other data cleaning approaches. Data cleaning involves approaches such as duplicate removal, compensation, and replacement. After data cleaning, the raw dataset is processed through imputations approaches. In this study, we imputed each case study dataset through imputation approaches such as the arithmetic mean, DNN, autoML, and an ensemble model. Hence, each imputation approach leads to generating a new imputed dataset.

The imputed datasets are validated using a deep learning-based regression model, and a dataset with improved performance results is selected as the final dataset for ceramic materials analysis applications. These applications include correlation analysis, input process and components parameters recommendation, and data augmentation and material discovery applications. In summary, we proposed an intelligent material data preparation mechanism for the AI-based ceramic material analysis application. The main contributions of the proposed study are as follows:

- 1) DNN-based imputation for missing values prediction.
- 2) AutoML-based imputation mechanism for automated material data pre-processing.
- 3) A novel approach to missing data imputation based on ensemble learning for data preparation and analysis.
- 4) Case study of novel PZT and TiO₃ ceramics materials as proof of concept.

The rest of the paper is organized as follows: A brief literature review is presented in Section 2. Section 3 presents the materials and methods for the proposed mechanism. Experimental results and performance analysis are discussed in Section 4. Finally, the conclusion and future work direction are presented in Section 5.

2. Related Work

Scholarly and industrial researchers have applied machine learning techniques for data analysis. Machine learning methods automate building analytical models based on algorithms learning from data.^[23] Classification and regression are areas in which machine learning has good applicability, and therefore it has played an important role in many fields.^[24–26] Materials science researchers have been using machine learning applications since the 1990s. For instance, the machine learning method was used for predicting ceramic matrix “composites” corrosion behavior and tensile strength.^[27,28] Therefore, machine learning has been used to study a wide range of topics in materials science, such as the discovery of new materials and the prediction of material properties.

Data pre-processing methods in datasets are divided into three core techniques: remove or drop the record, directly analyze missing values, and fill the missing data.^[10,29,30] Removing the missing values means removing the specific row or record where data is missing. This approach is best when the number of missing rows is less for the whole dataset. However, this technique will ignore some essential information and raise issues related to the analysis of the smaller datasets. Hence, the scientific community

has demonstrated the threat of completely deleting missing data using list wise removal approaches. Moreover, these deletion methods can present a considerable bias in the research, mostly when missing values are not distributed randomly. Thus the study concluded that this approach is defective in discrete datasets regarding classification accuracy.^[31] In this point of view, missing data filling is an area of stats and has engaged great attention in recent years.^[31,32]

Few missing data imputation approaches are based on the data mining probability. The filling of missing data starts from data analysis and is analyzed to impute the missing value based on likelihood knowledge. In literature, expectation maximization (EM) and linear regression (LR) are commonly used algorithms based on likelihood knowledge for data imputation. However, a poor understanding of the dataset distribution for EM and LR can predict dissimilarity. Hot-deck^[33,34] and multiple imputations are stats-based data imputation techniques adopted to develop scoring software to predict the mortality rate in ICU patients. The test result illustrates that the variations in ranges under the ROC curve were analytically essential but not logically suitable. For performance comparison, traditional imputation techniques were applied to different datasets to evaluate measures and analyze the differences.^[31] The results demonstrate that the hot deck is better than traditional imputation mechanisms. A study evaluated different imputation methods to overcome missing predictor values in a risk model.^[35] For instance, the performance of multivariate imputation (MI) and multiple imputation by chained equations (MICE) was better in terms of biased predictions and prognosis accuracy. Ultimately, Kenward and Carpenter^[36] summarize the impact of MI on contemporary software in health research. The researchers look at the problem of acquiring accurate imputations and spots that have proceeded as an output of the rising usage of MI, leading to future research guidance.

In the literature, learning models such as artificial neural network (ANN) is used to pre-process data, increase performance with respect to time, and reduce the noise produced due to the filling of missing values.^[37,38] Studies evidence that learning models perform better than traditional statistical techniques.^[39] Machine learning-based imputation methods are applied to generate predictive systems that predict missing values based on the dataset. In biomedical engineering fields such as bioinformatics, well-known learning algorithms such as multilayer perceptron (MLP), *k*-nearest neighbor (KNN), self-organizing maps (SOM), and decision tree (DT) have been used as imputation techniques. For instance, different algorithms were applied to the thyroid illness database, and some studies suggest that MLP-based imputation techniques perform better.^[40,41] Evaluated using current census data, MLP was also helpful in imputing various data values for review attributes.^[42] According to test evaluations, MLP consistently outperforms different approaches for imputing missing data.^[20,43] Batista et al. compared supervised learning with others missing data pre-processing techniques.^[44] The performance comparison results show that data filling approaches based on the KNN algorithm were better.^[45] Missing data from DNA microarrays have better imputation performance based on KNN.^[46,47] Similarly, SOM-based imputation methods have been applied to many problems, including the design of a real-size transport rating database and intelligent learning systems.^[30,48,49]

In summary, the scientific community believes encountering incomplete, missing, and insufficient data is common in today's data science applications. Imputation is the solution used to address missing or unknown data. In literature, researchers presented imputation mechanisms for filling the missing values with the mean, maximum, median, mode, or minimum values of the data characteristic. Some supervised learning and AI techniques were compared to the arithmetic mean, median, and mode. It is imperative to apply new intelligent techniques depending on the imputation target to not introduce bias into the data. Therefore, more sophisticated imputation techniques should be developed to fit the known data to a function or mathematical model based on advanced AI mechanisms.

3. Research Methodology

This section presents the methodology of the proposed intelligent material data preparation mechanism based on ensemble learning for AI-assisted material screening and discovery. The data imputation methodology is based on enhanced imputation mechanisms for generating imputed datasets. **Figure 2** presents the generic design of the proposed imputation methodology. First, a raw dataset is processed through imputations approaches based on statistical approaches and advanced machine learning techniques.

Then, for each imputation approach, an imputed dataset is produced. The imputed datasets are further validated using a regression model based on deep learning, and a dataset with improved performance results is selected as the final dataset for material analysis and other discovery applications. Similarly, the final prepared dataset is used for material applications such as correlation analysis, input process and properties recommendation, to name a few. **Figure 3** illustrates an example of the proposed methodology of material pre-processing and data preparation on PZT material dataset.

The first step is to create two instances of the dataset, one with missing values called raw data and the other with prepared data using traditional imputation techniques. The input to the pre-processing module is input composition data of the dataset, such as the composition of ZnO, MgO, and TiO₃, four process features, calcining temperature, calcining time, sintering temperature, and sintering time. The dataset also contains three output properties: sintered density, dielectric constant, and quality factor. The pre-processing module aims to find missing values and fill those using traditional approaches as well as proposed learning-based imputation techniques. Similarly, the proposed ensemble approach based on stacking is used to complete and impute the missing instances of the dataset. Stacking-based ensemble learning consolidates weak learners into a meta-learning model. This meta-learning model produces predictions based on the multiple predictions made by individual weak learning models. This study considers the model based on autoML and DNN as weak learners. Finally, the results of the learning models and traditional data cleaning approaches are compared in terms of regression performance analysis metrics. Similarly, the same methodology is applied to TiO₃ materials and it can be applied to any case study. For instance, in this study the TiO₃ materials input features are TiO₃, Ba, Bi, K, calcination temperature, calcination time, sintering temperature, and sintering time. The

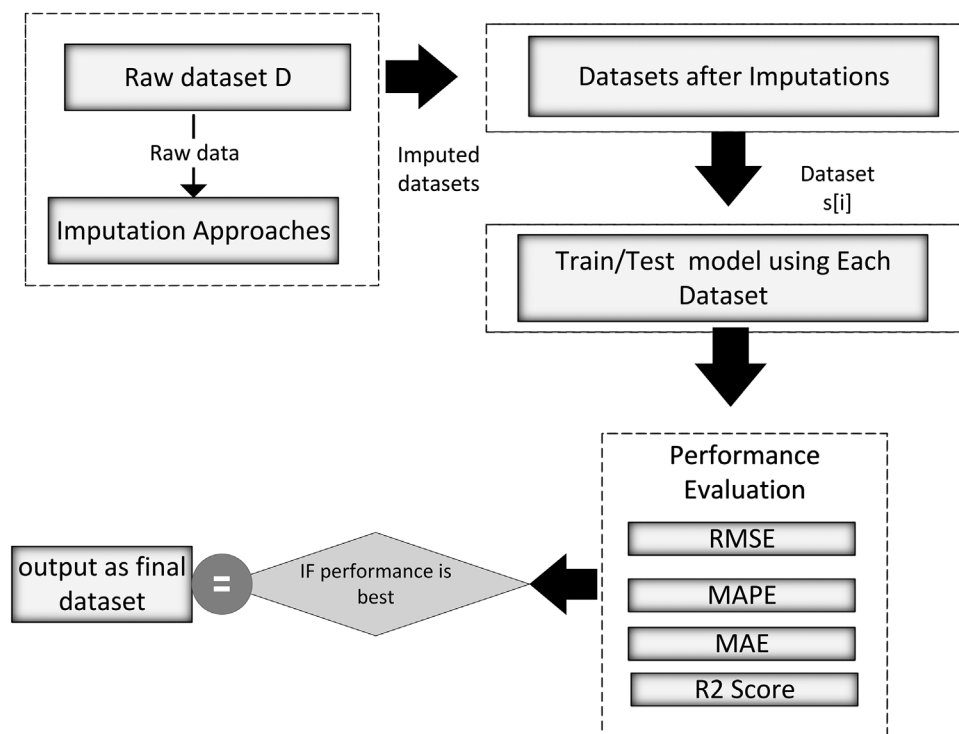


Figure 2. Proposed imputation mechanism for missing material data.

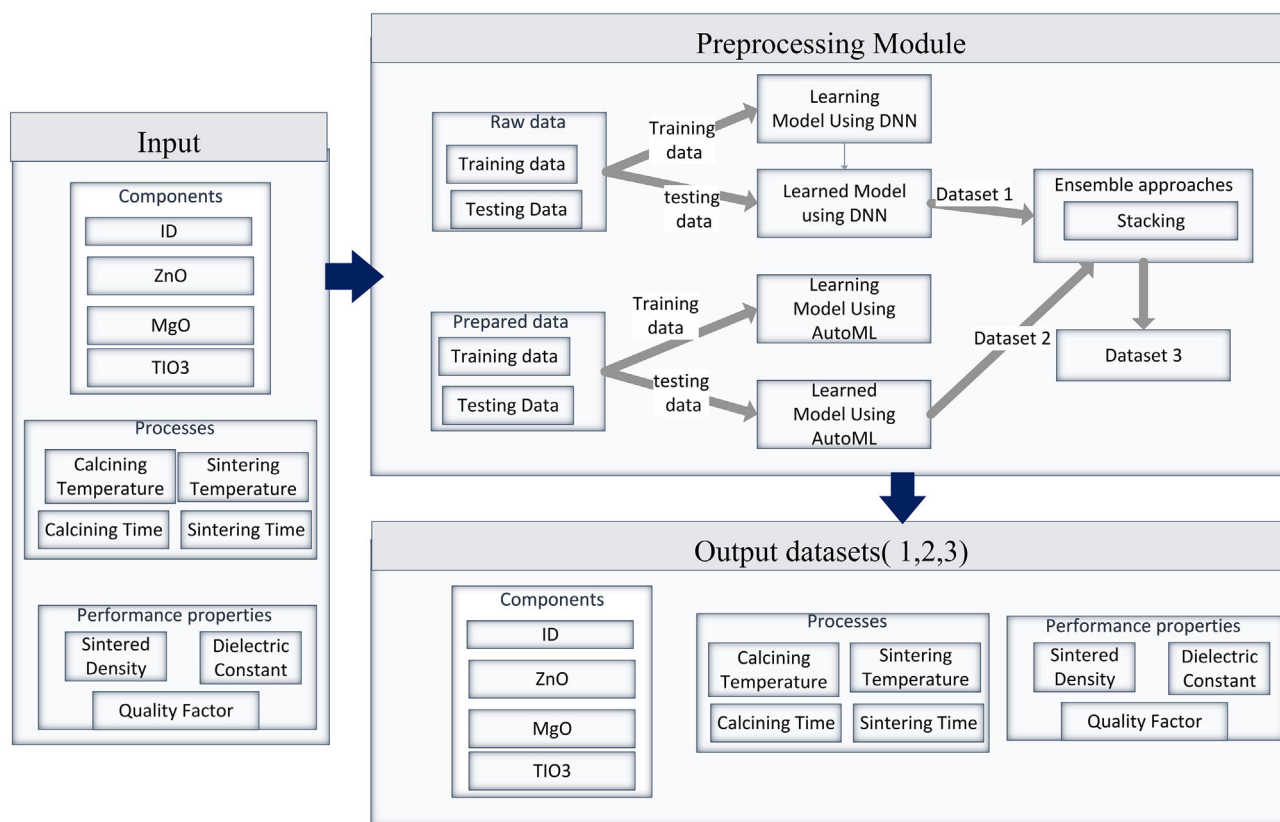


Figure 3. Proposed ensemble learning-based data imputation mechanism for PZT data.

Table 1. Configuration of the DNN for the imputation modeling and predictive analysis.

Input	Parameters	Output
Dataset 1		
i. Zn	i. Dense layers: 10	i. Sintered density
ii. MgO	ii. Training epochs:400	ii. Dielectric constant
iii. TiO ₃	iii. Batch size: 12	iii. Quality factor
iv. Calcining temperature	iv. Hidden layers activation function: ReLu	
v. Calcining time	v. Output layer activation function: linear	
vi. Sintering temperature	vi. Loss objective function: MSE	
vii. Sintering time		
Dataset 2		
i. Ba	i. Dense layers: 10	i. Temperature coefficient of resistivity
ii. Bi	ii. Training epochs:400	ii. Dielectric constant
iii. K	iii. Batch size: 12	iii. Electrical resistivity Curie temperature
iv. TiO ₃	iv. Hidden layers activation function: ReLu	
v. Calcining temperature	v. Output layer activation function: linear	
vi. Calcining time	vi. Loss objective function: MSE	
vii. Sintering temperature		
viii. Sintering time		

proposed “models” outputs are electrical resistivity, temperature coefficient of resistivity, Curie temperature, and dielectric constant.

3.1. Validation and Performance Analysis using Regression Model

This subsection presents the validation of the proposed research methodology using a deep learning-based regression model. As explained above, the proposed methodology outputs imputed and prepared datasets. Performance analysis in terms of regression metrics is used to verify and validate the datasets for machine learning-based material science applications. For the performance comparison, the regression model based on DNN is trained before and after the imputation of the case study datasets. A dataset with a low error rate and high R2 score is considered as reliable for machine learning models. **Table 1** presents the configuration of the DNN for the imputation modeling and predictive analysis. The raw data of the case study datasets consists of material components, processes, and raw performance index properties. The data cleaning approaches of missing values filling is performed to clean and prepare the dataset for recommendation models.

Case study dataset I, called PZT materials dataset, contains seven input features with three components: composition_MgO, composition_TiO₃, and four process features: calcining temperature, calcining time, sintering temperature, and sintering time. In addition, the PZT materials dataset contains three output performance index properties, property sintered density, property dielectric constant, and property quality factor, to validate the PZT materials. Similarly, the case study dataset II is called the TiO₃ materials dataset for the scope of this study. TiO₃ materials input features are TiO₃, Ba, Bi, K, calcination temperature, calcination time, sintering temperature, and sintering time. The proposed models’ outputs are electrical resistivity, temperature coefficient of resistivity, curie temperature, and dielectric constant.

4. Results and Discussion

This section presents the implementation results of the developed material analysis tool based on the proposed methodology. The proposed mechanisms were used to develop a material data preparation platform to provide a data cleaning mechanism based on identifying errors or corruptions, correcting or deleting them or manual processing data as needed to prevent the same errors from occurring. Most aspects of data cleaning can be done using the developed software tool based on the proposed methodology. Nevertheless, some manual inspection of the data and manual data preparation is still involved based on the Do It Yourself paradigm. Furthermore, it removes significant errors and inevitable inconsistencies when multiple data sources are pulled into one dataset. The platform cleans the raw dataset and exports the normalized dataset for other analysis applications.

4.1. Predictive Ceramics Analysis as Proof of Concept

Predictive analytics of ceramic material data is used to analyze the missing or incorrect material data using machine learning algorithms and statistical techniques to identify the likelihood of future outcomes based on current ceramic data. The goal is to go beyond what has happened to provide an enhanced assessment of what will happen if missing ceramic data is predicted. The pre-processing module aims to find missing values and fill them using traditional approaches and the proposed enhanced imputation mechanism. Ensemble approaches such as stacking and mean are used to complete the dataset. Finally, we compare all “imputation” approaches’ performance based on regression metrics.

The train-test split is a technique for data preparation for the evaluation of the performance of a machine learning algorithm. The procedure involves taking a dataset and dividing it into two subsets. The first subset is the training dataset, which is used to fit the model, and this process is called the model’s training. The second subset of the dataset is called the test dataset. Train subset is used to train the machine learning models such as autoML and

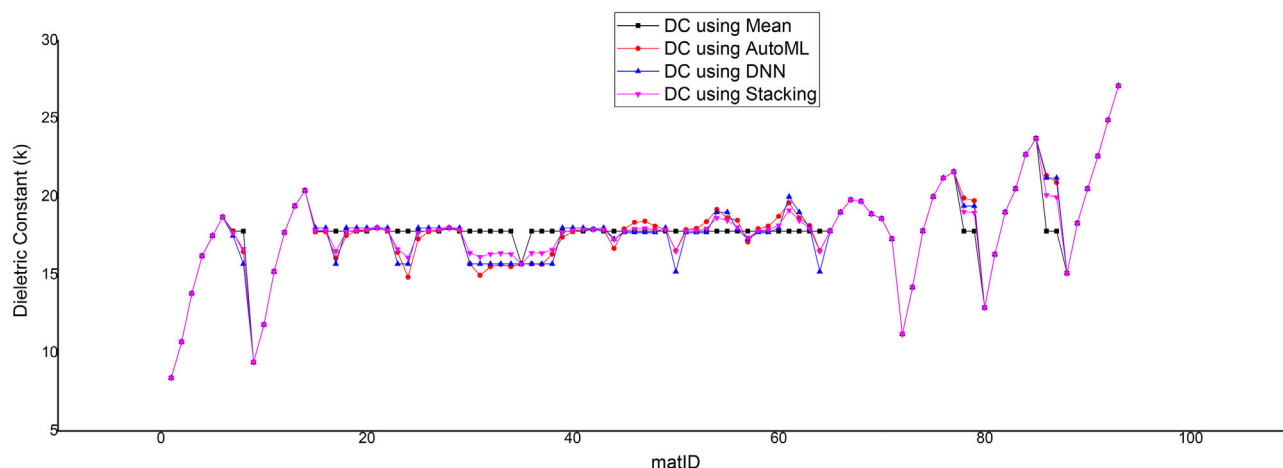


Figure 4. Imputation results' comparison of PZT- dielectric constant missing data.

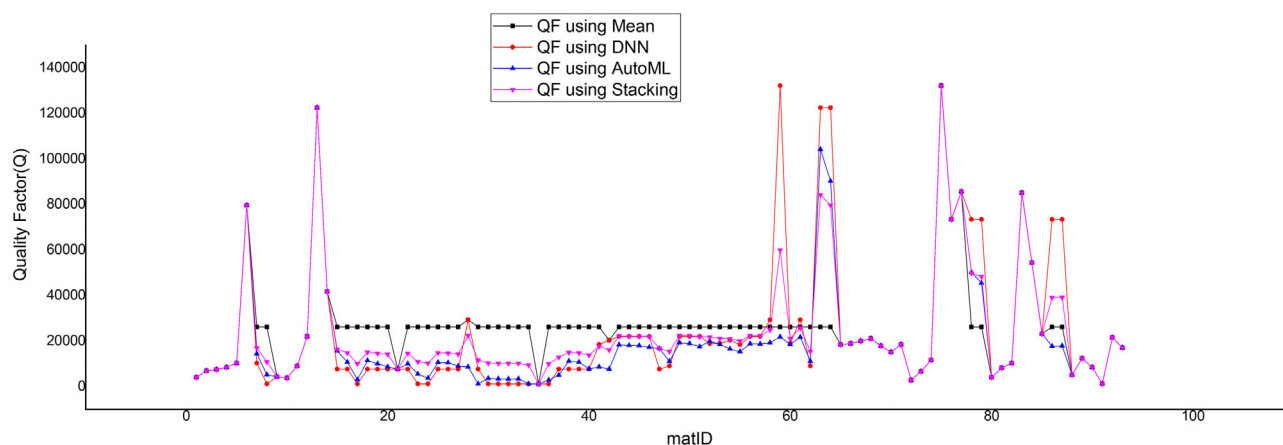


Figure 5. Imputation results' comparison of PZT- quality factor missing data.

DNN. The test subset is used to evaluate the inference on the unseen data. For instance, a training set of 0.67 percent means that the remaining percentage of 0.33 is assigned to the test set.

4.1.1. Predictive Analysis of PZT Ceramic Materials

This section compares the proposed enhanced imputation mechanism results with other imputation approaches based on PZT ceramic material data. **Figure 4** presents dielectric constant (DC) missing data prediction and comparison. The dielectric constant feature of the dataset is the dielectric constant property used for material performance evaluation. The x -axis represents the material ID, and the y -axis presents the DC values. The PZT material dataset used in this study needs imputation for missing DC and quality factors (QF) values. DC refers to the ratio of the substance's permittivity to the free-space permittivity. DC can be mathematically presented as

$$\kappa = \frac{\epsilon}{\epsilon_0} \quad (1)$$

The dielectric constant is essential for designing thin-film capacitors; in other circumstances, the material might be expected

to introduce capacitance into a circuit. Therefore, if the data is missing, the recommendation system will not work efficiently for the discovery and analytics functions. **Figure 5** shows a comparison of the DNN, autoML, stacking, and mean for missing QF values. The x -axis represents the material ID, and the y -axis presents the Q values. The quality factor performance property of the material is measured in unit Q , where Q represents the quality factor of the materials mixture system. We predict those missing values using the trained model of DNN, autoML, simple mean, and stacking approach.

4.1.2. Predictive Analysis of TiO₃ Materials

This section compares the proposed enhanced imputation mechanism results with other imputation approaches based on TiO₃ ceramic material data. TiO₃ stands for a combination of titanium (III) oxide in the material. The temperature coefficient of resistivity is the resistance-change factor per degree Celsius of temperature change. TiO₃ and Ba material dataset used in this study needs imputation for missing values of all data features. MatID is the ID of the material, and Ba component is Barium is a chemical element with the symbol Ba. **Figure 6** illustrates a compar-

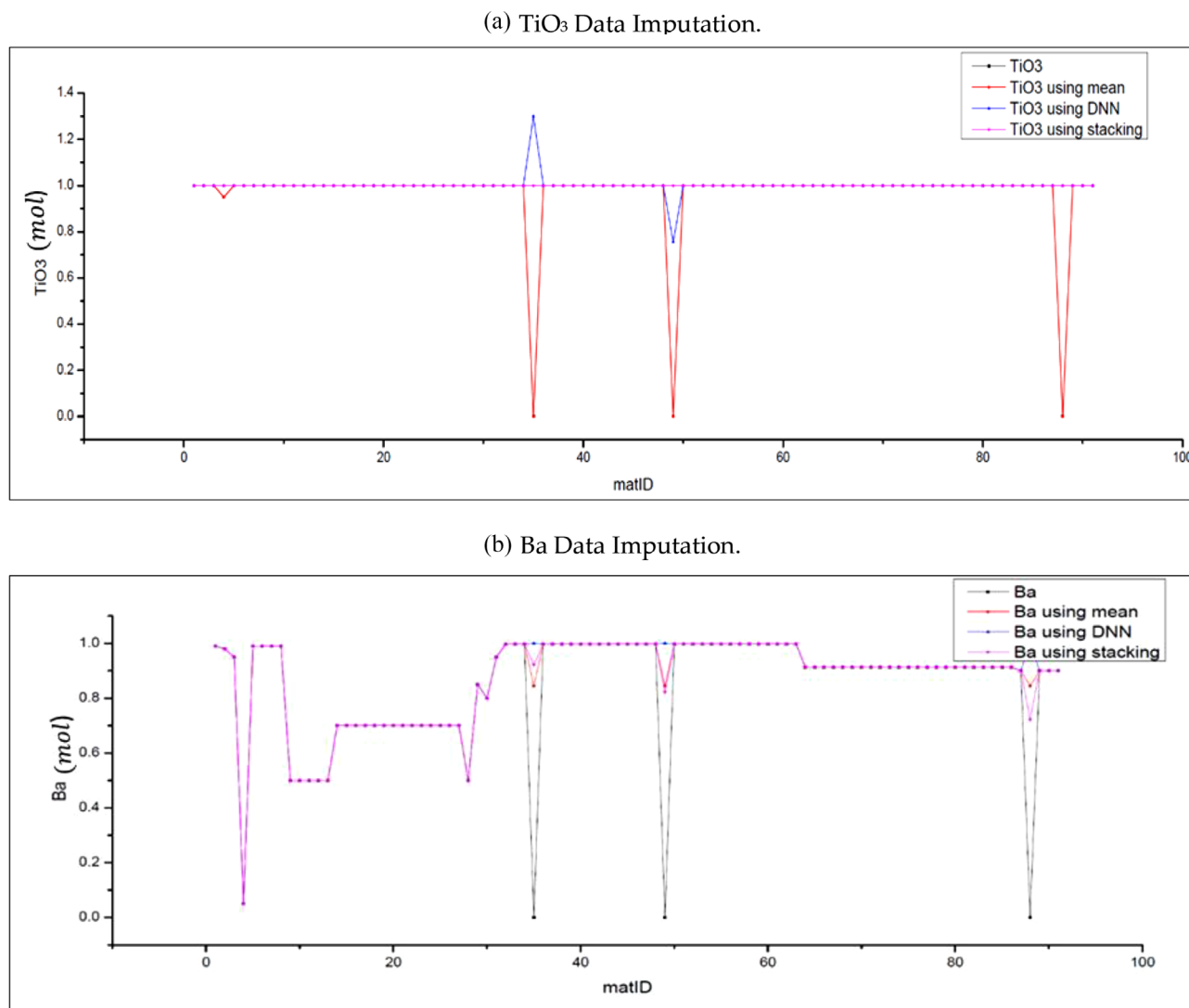


Figure 6. Comparison of missing values filling approaches for composition data. (a) TiO_3 missing values imputation. (b) Ba missing values imputation.

ison of the DNN, stacking, and mean-based approaches for Ba and TiO_3 values. The x -axis of Figure 6a,b represents the material ID (matID) of the material dataset instance. The y -axis of Figure 6a presents TiO_3 values and Figure 6b presents values of Ba.

Similarly, DNN, stacking, and mean-based approaches for Ba values are compared. This first step of data preparation for the imputation is filling the missing values with zero values. Then, we replaced zero values with mean values based on the zero-filled values to obtain the complete dataset with mean-based missing values imputation. The second dataset is obtained from the DNN predicted values, and the third dataset is obtained from the stacking-based approach. For instance, **Figure 7** presents a comparison of serial, stacking, and mean-based techniques for Bi values.

The x -axis of Figure 7 represents the material ID (matID) of the material dataset instance. Bi component data is data of Bismuth, a chemical element with the symbol Bi. For Bi components imputation, we first filled the missing values with zero values to

obtain our actual dataset. For the second dataset, the zero values of the Bi data feature were replaced with mean values. Similarly, the third dataset was obtained using DNN-based Bi values prediction. The last dataset for comparison was obtained using the proposed stacking approach. Finally, the model prediction using the model is compared to stacking and mean-based imputation. The calcination temperature process, calcination time in seconds, and sintering temperature data columns missing values imputations are presented in **Figure 8**.

The x -axis of Figure 8a–c represents the matID of the material dataset instance. We first filled the missing values with zero values; then, based on the zero-filled values, the mean-based imputation dataset was prepared. We predict those missing values using the learning models and finally impute using the ensemble model to make a comparison of these approaches. **Figure 9** compares serial, stacking, and mean-based approaches for electrical resistivity and coefficient of resistivity values. The SI unit of electrical resistivity is the ohmmeter, also denoted as $\Omega \cdot \text{m}$. The electrical resistivity is presented using Equation (2), where ρ_0

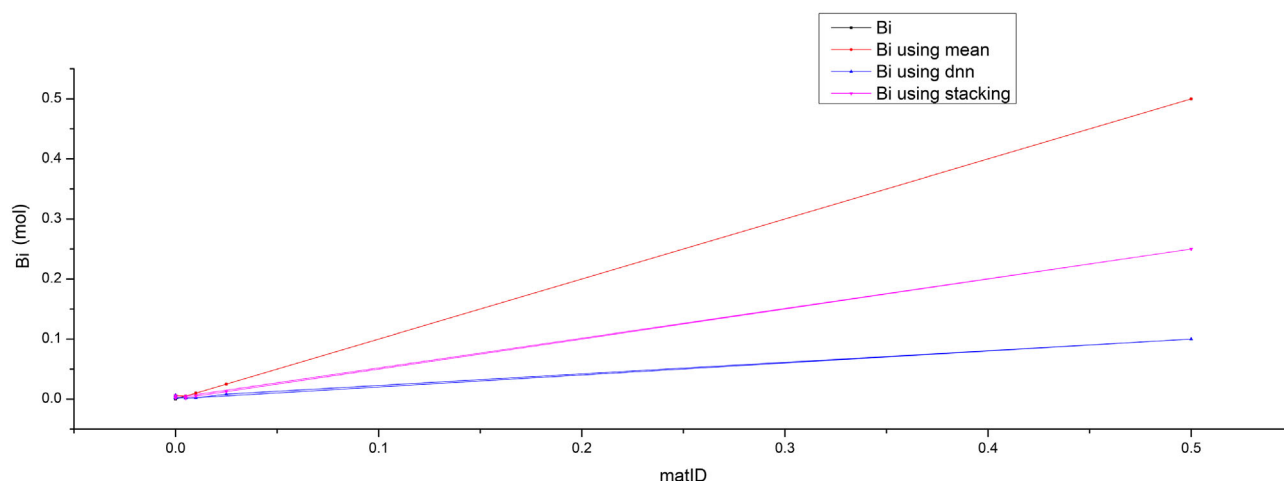


Figure 7. Comparison of imputation approaches for Bi missing values imputation.

represents the original electrical resistivity, and α is the coefficient of electrical resistivity.

$$\rho = \rho_0 (1 + \alpha \Delta T) \quad (2)$$

The x-axis of **Figure 10a,b** represents the matID of the material dataset instance. First, the dielectric constant missing values were replaced with zero values. Then, we replaced zero values with mean values to get one dataset based on the zero-filled values. Finally, we predict those missing values using the learned model of the serial approach and stacking approaches to perform a comparison. Figure 10 presents the comparison of imputation approaches for dielectric constant and Curie temperature values.

As the last step of the predictive analysis, we predicted the Curie temperature properties column contains missing values using the trained model. Similarly, using the trained model, the learning model predicts Curie temperature properties column contains missing values.

4.2. Performance Analysis

This section presents the performance analysis of the proposed imputations mechanisms. To begin with, we will discuss the performance evaluation metrics used for the case studies results evaluation. The performance evaluation metrics used for this include mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and R-squared or coefficient of determination metrics are used to evaluate the model's performance in learning models of the recommendation system. Mean absolute error is utilized to evaluate PZT and TiO₃ materials recommendations to measure the deviation between actual PZT and TiO₃ values and imputed values. It is obtained by taking the difference between the inferred and target values as given by

$$MAE = \frac{\sum_{i=1}^n |Y_i - \hat{Y}_i|}{n} \quad (3)$$

Mean squared error measures the average of the squared difference between the inferred and actual PZT and TiO₃ values; the square is taken to remove negative values; the following equation gives MSE

$$MSE = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n} \quad (4)$$

Root mean squared error is utilized to discover the error rate of the learning model and check the error size to be the same as the size of target PZT and TiO₃ values. It is determined by taking the square root of MSE

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}} \quad (5)$$

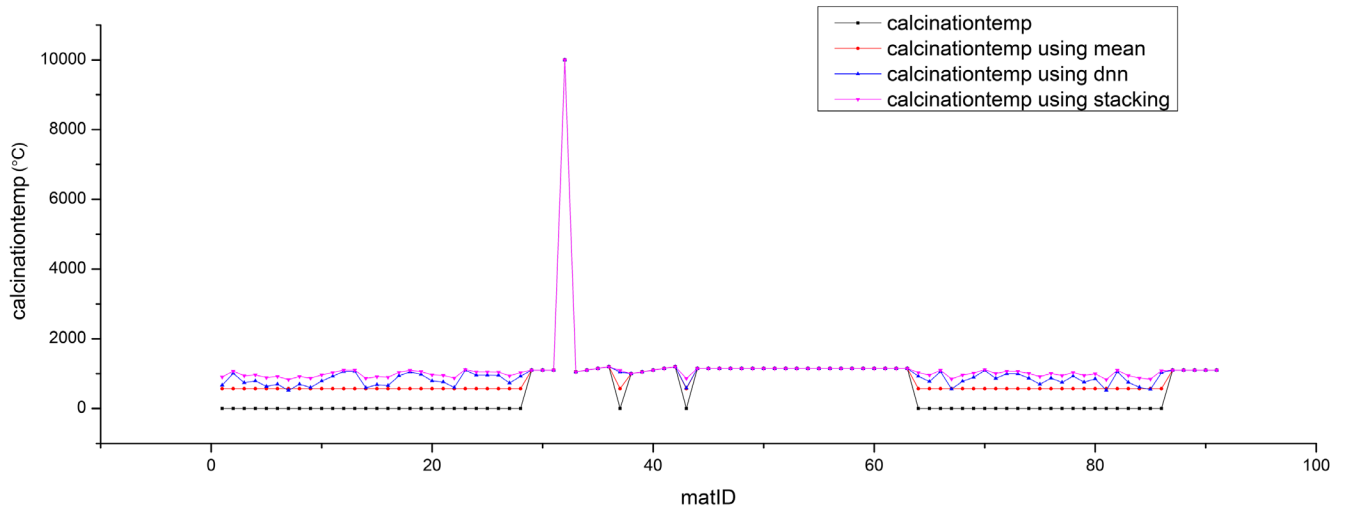
Mean absolute percentage error is the average of absolute errors divided by the original values and multiplied by 100 to get the result in a percentage.

$$MAPE = \frac{\sum_{i=1}^n \frac{|Y_i - \hat{Y}_i|}{Y_i}}{n} \times 100 \quad (6)$$

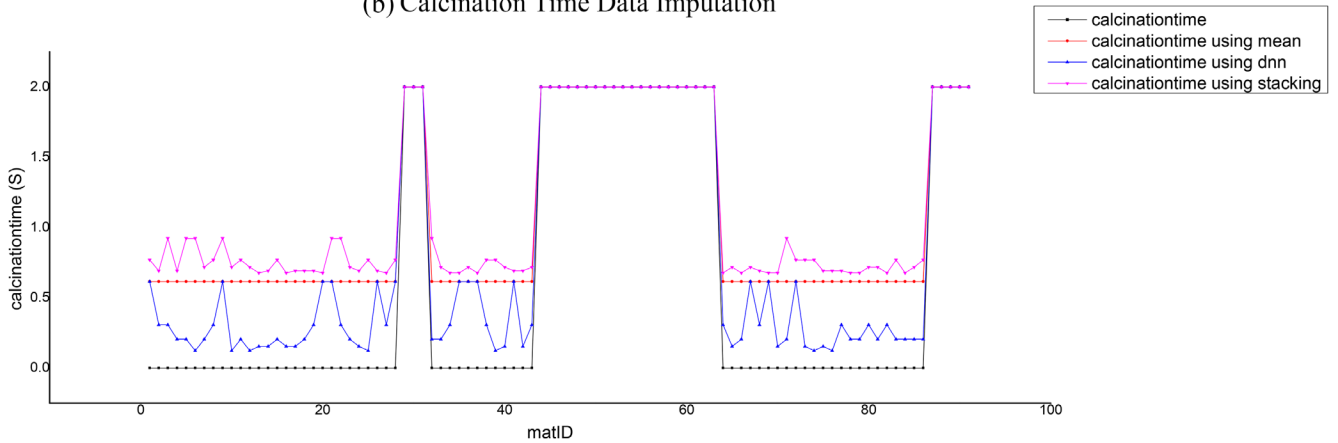
RMSE is more extensively used than MSE to evaluate the performance of the learning models. Therefore, RMSE is used as a primary loss function in the learning models despite being more complicated to interpret than MAE; for instance, MAE is better for data containing outliers. The lower value of MAE, MSE, and RMSE indicates the higher accuracy of the learning model. A greater value of R square is regarded as excellent. R Squared and adjusted R-squared are utilized to demonstrate how well the independent features explain the variability in the dependent data features.

Table 2 presents a comparative analysis of imputation approaches to the PZT dataset. The learning model's performance is evaluated in terms of MSE, RMSE, and R2 score. The results are compared using four datasets achieved through four data pre-processing methods: mean approach, DNN, autoML, and

(a) Calcination Temperature Data Imputation



(b) Calcination Time Data Imputation



(c) Sintering temperature data Imputation

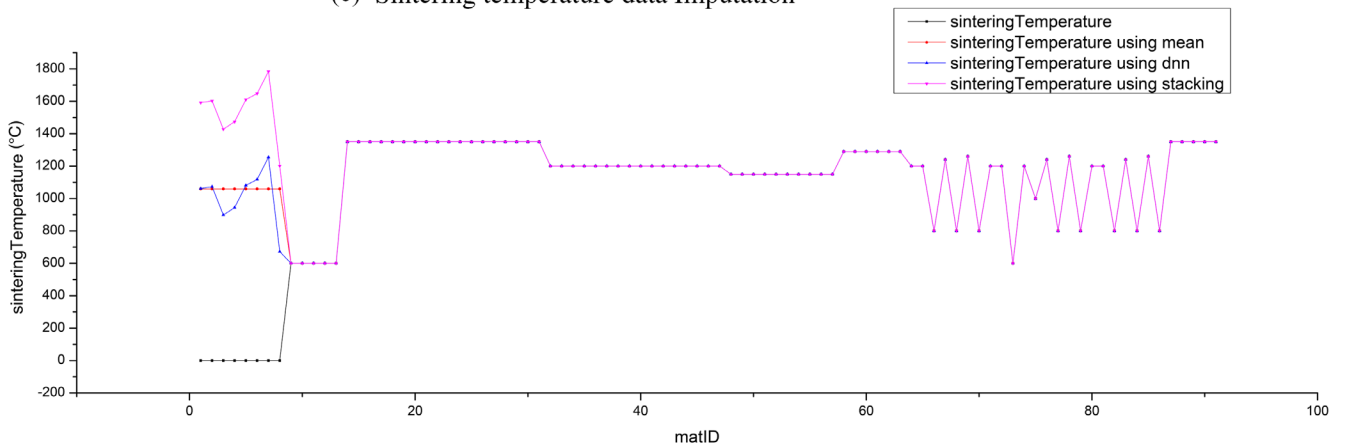


Figure 8. Comparison of imputation approaches for processes parameters values imputation. (a) calcination temperature missing values imputation. (b) calcination time missing values imputation. (c) sintering temperature missing values imputation.

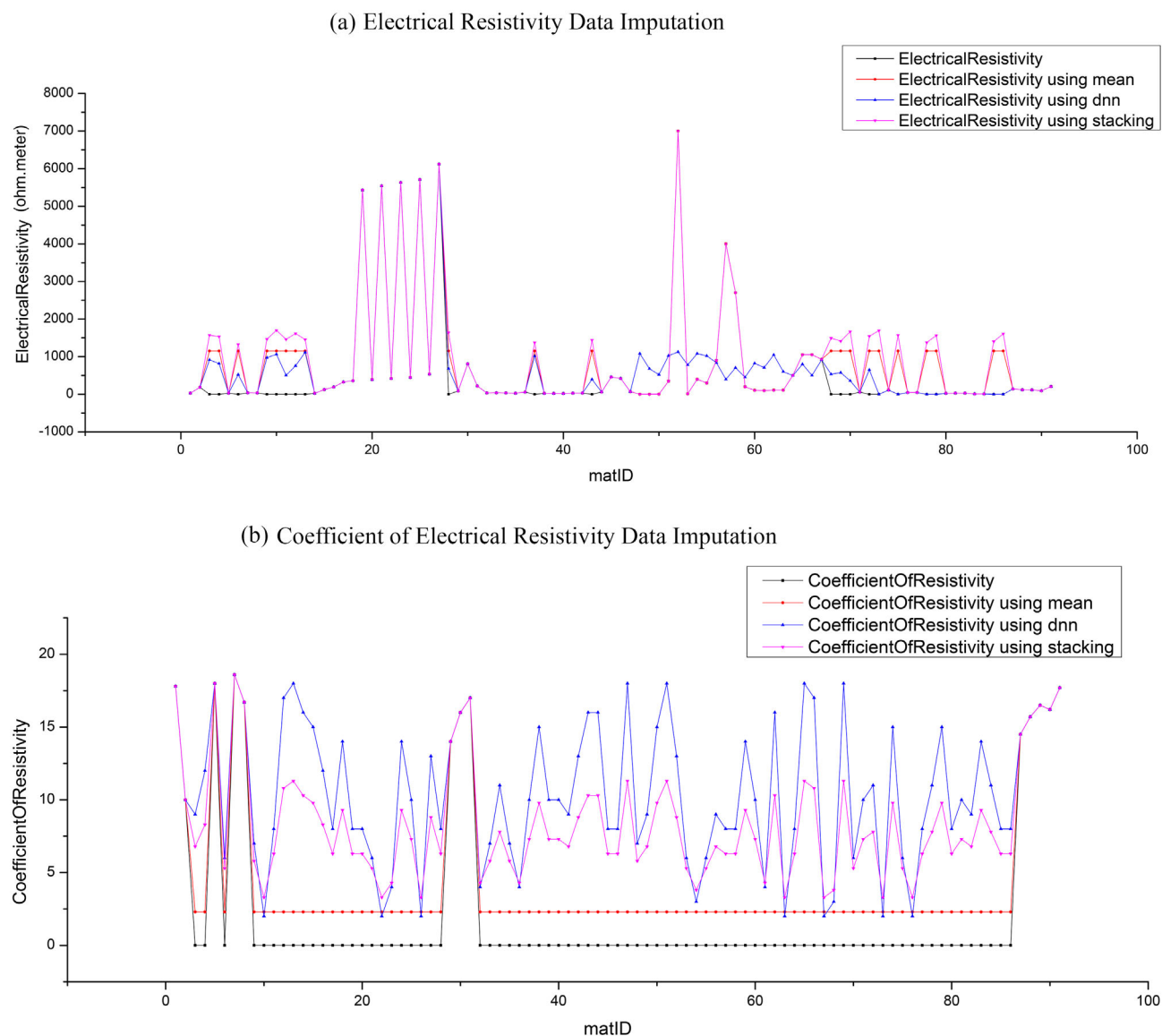


Figure 9. Comparison of imputation approaches for electrical resistivity and coefficient of resistivity missing values imputation. (a) Electrical resistivity missing values imputation. (b) Coefficient of resistivity missing values imputation.

stacking approach. Results indicate that the learning model trained based on the proposed stacking approach generated dataset performs better in terms of RMSE of 49.18 and an R2 score of 0.59. Moreover, in the traditional mean-based filling of missing values, the model performance in terms of R2 score was negative. The R2 value is negative if the chosen model does not follow the trend of the data that is, it fits worse than a horizontal line. This means that when the missing data is filled through the traditional mean approach, the model cannot follow the trends of material data. However, when the missing data records were filled using our proposed approaches, the R2 value was improved.

Similarly, Table 3 presents a comparative analysis of imputation approaches based on the TiO₃ dataset. The results are compared using four datasets achieved through three data pre-processing techniques: mean approach, DNN, autoML, and

stacking approach. Results indicate that the proposed learning models trained based on the stacking approach generated dataset perform better with RMSE of 97.97, MAPE of 68.051, and R2 score of 0.76.

5. Conclusions

The industrial applications of ceramics with special experimental conditions and limited resources require new scientific practices and applications. As a result, enormous material data is being collected from traditional and high-tech experiments. However, the manual recording reveals inconsistencies in the analysis results. Therefore, complete and reliable material data is vital for advanced AI applications such as descriptive and predictive analysis and optimization of components and process parameters

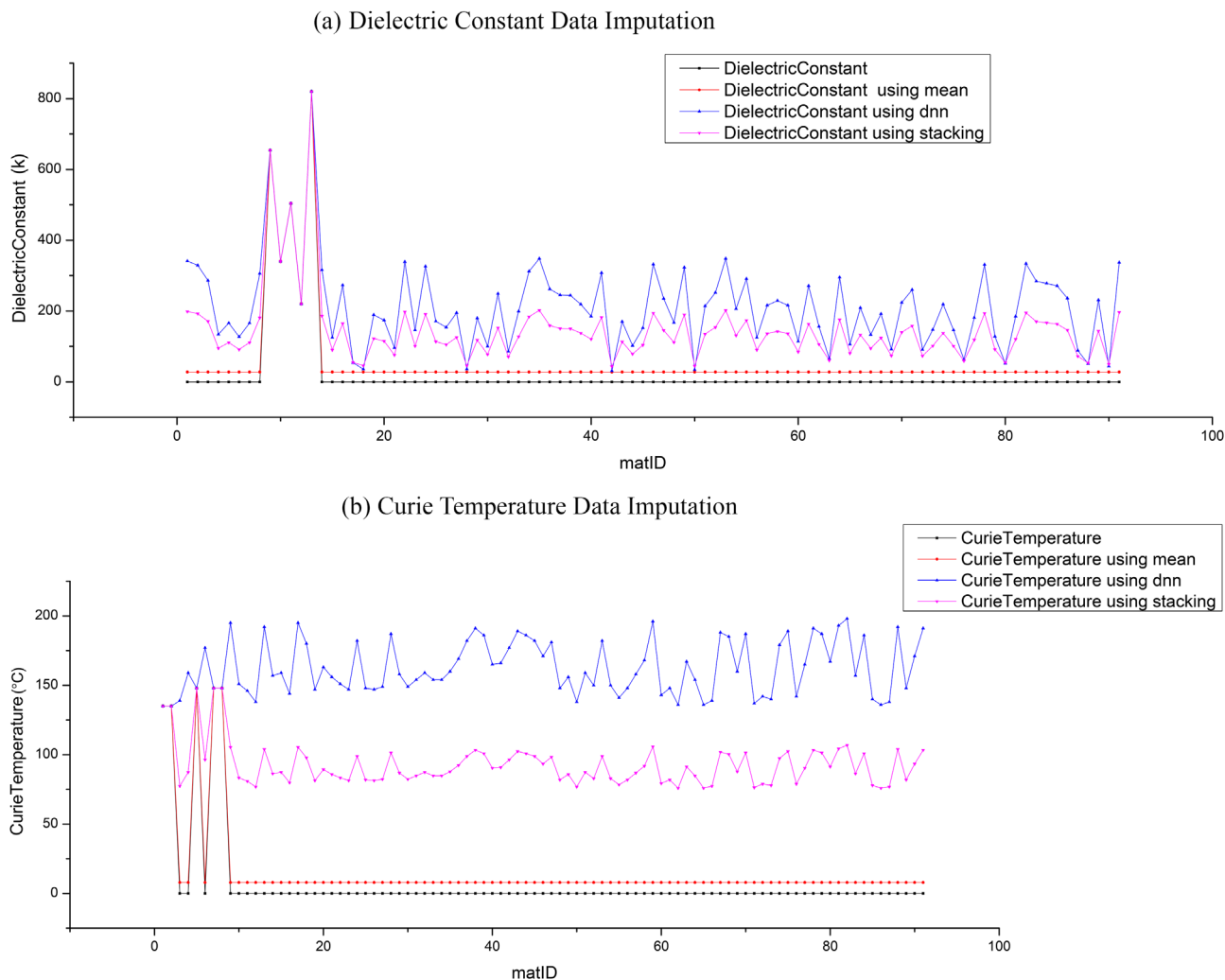


Figure 10. Comparison of imputation approaches for performance index missing values imputation. (a) Dielectric constant missing values imputation. (b) Curie temperature missing values imputation.

Table 2. Comparative analysis of imputation approaches on PZT dataset.

Performance metric	Mean approach	DNN approach	AutoML approach	Ensemble approach
MSE	4433.76	5239.35	5148.73	2419.31
RMSE	66.58	72.38	71.75	49.18
R2 score	−0.030	0.049	0.40	0.59

Table 3. Comparative analysis of imputation approaches based on the TiO₃ dataset.

Performance metric	Mean approach	DNN approach	Ensemble approach
MAPE	136.62	102.33	68.05
MSE	31825.92	20712	9598.36
RMSE	180.47	143.91	97.97
R2 score	0.21	0.48	0.76

to discover better material. Thus, intelligent approaches for imputing missing material data are needed to minimize the impact of incomplete and missing data for AI applications. In the current study, we proposed a novel material pre-processing and data preparation mechanism based on an intelligent imputation methodology based on ensemble learning of deep learning and automated machine learning models. The proposed methodology prepare and impute the data using traditional mean and advanced machine learning models such as autoML, DNN, and an ensemble model resulting in four datasets for an input dataset.

As a case study, two ceramics datasets, are used to validate and verify the proposed methodology. The performance analysis is evaluated using regression performance matrices RMSE, MAE, MAPE, and R2 score. Thus deep learning-based regression model is trained before and after the imputation techniques applied to the datasets. This comparison is made to prove the effectiveness of the proposed ensemble model for data preparation for material analysis and recommendation applications in terms of

error rate. Here is a summary of the findings from this research study:

- 1) Machine learning models trained with datasets imputed through traditional mean and average-based approaches do not follow the data trends and thus lead to negative R2 scores.
- 2) The dataset obtained and prepared through a learning-based mechanism is better and more reliable for learning model-based material applications.
- 3) Though learning mechanisms such as DNN and autoML can be reliable instead of traditional statistical approaches, the results can be enhanced further using ensemble models.
- 4) In the case study of the PZT dataset, the performance analysis shows that the learning model trained based on the proposed ensemble model-based generated dataset performs better in terms of RMSE of 49.18 and an R2 score of 0.59.
- 5) Similarly, in the case of the TiO₃ dataset, the ensemble model performs better with an RMSE of 97.97, MAPE of 68.051, and R2 score of 0.76.
- 6) With traditional mean-based missing-value filling, the model performance leads to a negative value of the R2 score. This means that if the traditional mean approach fills in the missing data, the model cannot follow the trends in the material data.

As part of future work, the proposed mechanism can be used to solve the problem of smaller datasets using data augmentation mechanisms. Furthermore, this study provides promising directions to develop AI-based material discovery and analysis applications based on the prepared datasets using the proposed methodology. As a future study, we will design and develop a material assessment and discovery framework for advanced AI-based material discovery and recommendation.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

ceramic data pre-processing, ceramic material, data imputation, material analysis, material data preparation

Received: July 19, 2022
Revised: August 25, 2022
Published online:

- [1] N. Iqbal, A.-N. Khan, Imran, A. Rizwan, F. Qayyum, S. Malik, R. Ahmad, D.-H. Kim, *J. Manuf. Syst.* **2022**, 64, 19.
- [2] M. Imran, U. Zaman, Imran, J. Imtiaz, M. Fayaz, J. Gwak, *Electronics* **2021**, 10, 2501.

- [3] I. Imran, U. Zaman, M. Waqar, A. Zaman, *Soft Comput. Mach. Intell.* **2021**, 1, 11.
- [4] Imran, F. Jamil, D. Kim, *Sustainability* **2021**, 13, 10057.
- [5] Imran, D. H. Kim, *J. Intell. Pervasive Soft Comput.* **2022**, 1, 10.
- [6] Imran, F. Qayyum, D.-H. Kim, S.-J. Bong, S.-Y. Chi, Y.-H. Choi, *Materials* **2022**, 15, 1428.
- [7] M. T. Alam, *Mater. Sci. Forum* **2021**, 1034, 73.
- [8] P. D. Allison, *Missing Data*, Sage Publications, Thousand Oaks, CA **2001**.
- [9] R. J. A. Little, D. B. Rubin, *Statistical Analysis with Missing*, John Wiley & Sons, Inc, Hoboken, NJ **2002**.
- [10] J. G. Ibrahim, M. H. Chen, S. R. Lipsitz, A. H. Herring, *J. Am. Stat. Assoc.* **2005**, 100, 332.
- [11] A. Gupta, M. S. Lam, *J. Oper. Res. Soc.* **1996**, 47, 229.
- [12] P. K. Sharpe, R. J. Solly, *Neural Comput. Appl.* **1995**, 3, 73.
- [13] G. Ambler, R. Z. Omar, P. Royston, *Stat. Methods Med. Res.* **2007**, 16, 277.
- [14] B. G. EAPA, M. C. Monard, in *Int. Conf. on Hybrid Intelligent Systems* (Eds.: J. R. S. Abraham, M. Köppen), DBLP, Santiago, Chile **2002**, pp. 251–260.
- [15] E. R. Hruschka, E. R. Hruschka, N. F. F. Ebecken, in *AI 2004: Advances in Artificial Intelligence*, Springer, Berlin/Heidelberg **2005**, p. 513.
- [16] P. J. G. Lisboa, H. Wong, P. Harris, R. Swindell, *Artif. Intell. Med.* **2003**, 28, 1.
- [17] J. L. Schafer, J. W. Graham, **2002**, 7, 147.
- [18] G. Hawthorne, G. Hawthorne, P. Elliott, *Aust. N. Z. J. Psychiatry* **2005**, 39, 583.
- [19] E. Biganzoli, P. Boracchi, L. Mariani, E. Marubini, *Stat. Med.* **1998**, 17, 1169.
- [20] B. D. Ripley, R. M. Ripley, in *Clinical Applications of Artificial Neural Networks* (Eds: R. Dybowski, V. Gant), Cambridge University Press, Cambridge **2001**, p. 237.
- [21] Z. Zheng, Y. Qu, W. Ma, F. Hou, *Acta Mater. Compositae Sin.* **1998**, 15, 14.
- [22] T. Furukawa, K. Ishida, E. Fukada, *J. Appl. Phys.* **1979**, 50, 4904.
- [23] Imran, N. I., D. H. Kim, *Energy Build.* **2022**, 257, 111762.
- [24] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, V. Robles, *Briefings Bioinf.* **2006**, 7, 86.
- [25] M. Eminagaoglu, S. Eren, in *2010 Int. Conf. on Computer Information Systems and Industrial Management Applications (CISIM)*, IEEE, Piscataway, NJ **2010**, pp. 187.
- [26] Y. Kang, Z. Cai, C.-W. Tan, Q. Huang, H. Liu, *J. Manage. Anal.* **2020**, 7, 139.
- [27] L. Chonghe, G. Jin, Q. Pei, C. Ruiliang, C. Nianyi, *J. Phys. Chem. Solids* **1996**, 57, 1797.
- [28] H. S. Rao, A. Mukherjee, *Comput. Mater. Sci.* **1996**, 5, 307.
- [29] D. V. Nguyen, N. Wang, R. J. Carroll, *Data Sci.* **2004**, 2, 347.
- [30] K. Lakshminarayan, S. A. Harp, T. Samad, *Intelligence* **1999**, 11, 259.
- [31] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Inc, Hoboken, NJ **1987**.
- [32] C. F. Manski, *Int. J. Approx. Reasoning* **2005**, 39, 151.
- [33] A. Pérez, R. J. Dennis, J. F. A. Gil, M. A. Rondón, A. López, *Stat. Med.* **2002**, 21, 3885.
- [34] J. Siddique, T. R., *Medicine (Baltimore)* **2008**, 27, 83.
- [35] J. M. Jerez, L. Franco, E. Alba, A. Llombart-Cussac, A. Lluch, N. Ribelles, B. Munárriz, M. Martín, *Breast Cancer Res. Treat.* **2005**, 94, 265.
- [36] M. G. Kenward, J. Carpenter, *Stat. Methods Med. Res.* **2007**, 16, 199.
- [37] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell, *Cancer* **1997**, 79, 857.
- [38] K. Fujikawa, Y. Matsui, T. Kobayashi, K. Miura, H. Oka, S. Fukuzawa, *J. Urol.* **2003**, 10, 149.

- [39] W. Duch, J. Kacprzyk, E. Oja, S. Zadrozny, in *Artificial Neural Networks: Biological Inspirations ICANN 2005: 15th Int. Conf. Warsaw*, Springer, Berlin, Germany **2005**, p. 15.
- [40] A. M. Mendoza, R. M. Hernandez, in *13th Int. Conf. on Information and Communication Technology and System*, IEEE, Piscataway, NJ **2021**, pp. 207.
- [41] F. Temurtas, *Expert Syst. Appl.* **2009**, 36, 944.
- [42] N. Svein, *Off. Stat.* **1996**, 12, 385.
- [43] D. A. Elizondo, M. A. Gongora, *ICANN, International conference on Artificial neural network*, Springer, Warsaw, Poland **2016**, pp. 485–490.
- [44] P. Clark, R. Boswell, in *Machine Learning - EWSL-91 : European Working Session on Learning* (Ed.: Y. Kodratoff), Springer, Berlin, Heidelberg **1991**, pp. 151–163.
- [45] G. E. A. P. A. Batista, M. C. Monard, *Appl. Artif. Intell.* **2003**, 17, 519.
- [46] I. Scheel, M. Aldrin, I. K. Glad, R. Shrum, H. Lyng, A. Frigessi, **2005**, 21, 4272.
- [47] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. B. Altman, **2001**, 17, 520.
- [48] F. Fessant, S. Midenet, *Neural Comput. Appl.* **2002**, 10, 300.
- [49] T. A. Md, *Material science Forum* **2002**, 1034, 73.