ACCSCIENCE PUBLISHING

Materials Science in Additive Manufacturing

ORIGINAL RESEARCH ARTICLE

Data imputation strategies for process optimization of laser powder bed fusion of Ti6Al4V using machine learning

Guo Dong Goh[1], Xi Huang[2], Sheng Huang[3], Jia Li Janessa Thong[3], Jia Jun Seah[3], Wai Yee Yeong[1,2,3]*

[1]Singapore Centre for 3D Printing, School of Mechanical and Aerospace Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore

[2]HP-NTU Digital Manufacturing Corporate Lab, Nanyang Technological University, Singapore

[3]School of Mechanical and Aerospace Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore

Abstract

A database linking process parameters and material properties for additive manufacturing enables the performance of the material to be determined based on the process parameters, which are useful in the design and fabrication stage of a product. The data, however, are often incomplete as each individual research work focused on certain process parameters and material properties due to the wide range of variables available. Imputation of missing data is thus required to complete the material library. In this work, we attempt to collate the data of Ti6Al4V, a popular alloy used in aerospace and biomedical industries, fabricated using powder bed fusion, or commonly known as selective laser melting (SLM). Various imputation techniques of missing data of the SLM Ti6Al4V dataset, such as the k-nearest neighbor (kNN), multivariate imputation by chained equations, and graph imputation neural network (GINN) are investigated in this article. It was observed that kNN performed better in imputing variables related to process parameters, whereas GINN performed better in variables related to material properties. To further improve the quality of imputation, a strategy to use the median of the imputed values obtained from the three models has resulted in significant improvement in terms of the relative mean square error. Self-organizing map was used to visualize the relationship among the process parameters and the material properties.

Keywords: Additive manufacturing; 3D printing; Selective laser melting; Powder bed fusion; Machine learning; Data analytics; Imputation

***Corresponding author:**
Wai Yee Yeong
(wyyeong@ntu.edu.sg)

Citation: Goh GD, Huang X, Huang S, *et al.*, 2023, Data imputation strategies for process optimization of laser powder bed fusion of Ti6Al4V using machine learning. *Mater Sci Add Manuf,* 2(1): 50.
https://doi.org/10.36922/msam.50

Received: February 1, 2023

Accepted: March 7, 2023

Published Online: March 22, 2023

1. Introduction

Ti6Al4V is one of the most popular titanium alloys given its excellent material properties, including high strength, low density, and high corrosion resistance, and is used in a wide variety of industries, such as in aerospace for aircraft components and in biomedical for implants[1]. Instead of using traditional manufacturing methods, selective laser melting (SLM) of Ti6Al4V allows for more complex parts to be created. It is an additive manufacturing technique, categorized as powder bed fusion (PBF), which involves

melting layers of powder to form functional parts using computer-aided design (CAD) software which allows for freedom of design[2]. The SLM process is influenced by a set of process parameters; however, the number of these parameters is not clearly defined. The most influential ones are shown in Figure 1.

Typically for a material to be established as processable by SLM, parametric studies are needed to optimize the parameters to obtain defect-free and fully dense parts. These parameters may differ from one machine to another and between materials. The common parameters studied are the laser power, laser scanning speed, hatch spacing, and powder layer thickness. In particular, the laser power controls the amount of energy that irradiates the material while hatch spacing which defines the distance between two laser scans and it should provide enough overlap between adjacent scan tracks to bonds. Controlling the bed density in SLM can be challenging as it depends on several factors, such as the powder size and shape, the recoating process, and the initial bed leveling. The bed density can affect the heat transfer during the melting process and the resulting microstructure and mechanical properties of the printed parts.

One of the attractive points of SLM Ti6Al4V is that the material properties can be tuned by optimizing the SLM process parameters. For instance, by adjusting the scanning speed and hatching distance, Roudnicka *et al.* investigated various energy density values ranging from 40 to 400 J/mm^3[3]. Their results showed that porosity and mechanical properties can be significantly altered by adjusting the parameters, and suggested a processing range for achieving the highest relative density. Furthermore, it was found that modifying the energy density may cause microstructural changes, which can affect the mechanical properties of the final product. Different processing conditions will result in different microstructures and mechanical properties. For instance, the microstructure of SLM-produced bulk samples consists of needle-shaped α'-martensite phase[4]. Heat treatment promotes the formation of free α- and β-phases at grain boundaries. The samples possess high strength but low ductility prior to heat treatment. Heat treatment enhances their mechanical properties at both ambient and elevated temperatures. The previous studies have researched the process-properties relationships for SLM Ti6Al4V but have focused only on a few properties in their studies: physical properties such as relative density[5-9], or mechanical properties such as tensile strength and Young's modulus[10-14].

Other studies that have carried out parameter optimization or modeling for Ti6Al4V do not investigate the full range of process parameters and only involve a few material properties. Sun *et al.*[15] and Kuo *et al.*[6] have investigated the relationship between process parameters and the density of the SLM Ti6Al4V. The former varied laser power, scanning speed, layer thickness, hatch spacing, and scanning strategy to obtain specimens with maximum density, while the latter varied laser power, exposure duration, and point distance and reported the porosity of the printed specimen. Bartolomeu *et al.*[16] studied and modeled the effects of laser power, scan speed, and hatch spacing on the density, hardness, and shear strength of SLM Ti6Al4V. More research have been conducted to understand the fatigue behavior of SLM-fabricated Ti-6Al-4V[17], which may differ from the behaviors under static loading.

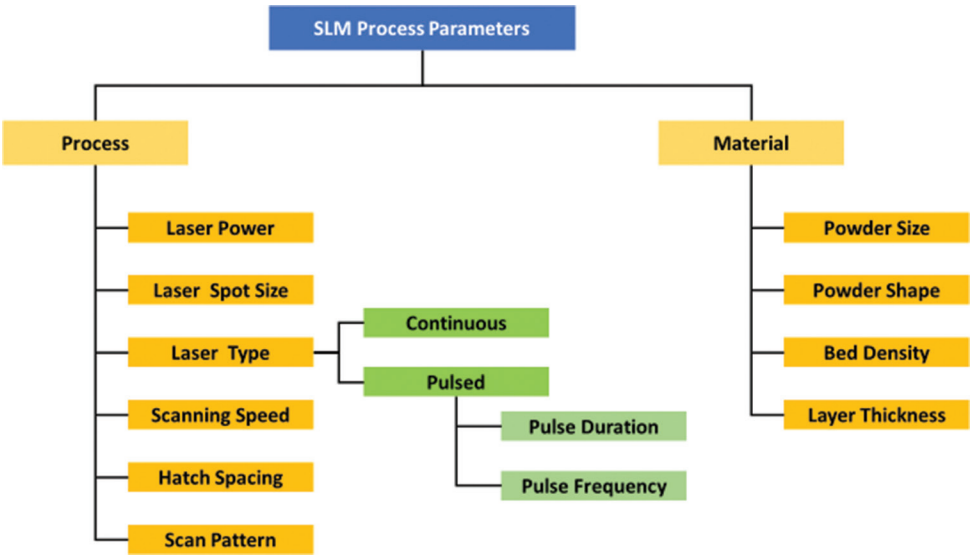

Figure 1. Influential parameters that affect the quality of the part fabricated by selective laser melting.

Machine learning in 3D printing is growing rapidly and has been used to perform process and design optimization, anomalies detection, etc.[18]. It relies heavily on the dataset to train a good machine learning model to have good prediction. Given the vast number of literature investigating the process parameters' effects on the different properties of SLM Ti6Al4V, there is potential in collating the data and using machine learning to perform data analytics on the dataset to determine the process-structure-properties relationship. There are missing values present in the collated SLM Ti6Al4V dataset as each property/parameter has been studied in isolation, but the quantity of data is insufficient for machine learning; therefore, imputation is required to bolster the data volume. Hence, the data from the literature are considered incomplete, and imputation of the missing data is required as a pre-processing step before subsequent analysis can be carried out.

Researchers have utilized various kinds of techniques to impute missing data in manufacturing processes. For instance, Steiner *et al.* aimed to develop real-time predictive models of two key strength properties of a wood composite manufacturing process using real-time process and destructive test data collected from a wood composite manufacturer[19]. However, sensor malfunction and data "send/retrieval" problems lead to null fields in the company's data warehouse, which resulted in information loss. To overcome this challenge, two missing data imputation methods, expectation-maximization (EM) algorithm and multiple imputation (MI) using Markov Chain Monte Carlo (MCMC) simulation, were used to impute the missing data. Predictive models based on the imputed datasets generated more precise prediction results than models of non-imputed datasets. In addition, Bayesian Additive Regression Tree (BART) produced the most precise prediction results among four predictive modeling methods. In another work, Wang *et al.* discuss the importance of data mining in intelligent manufacturing and introduce an energy monitoring platform for small- and medium-sized enterprises that records energy consumption data at various levels of granularity[20]. However, incomplete data can lead to an inaccurate portrayal of the system, so Wang *et al.* propose a novel orthogonal-least-square-based autoencoder to generate new samples for the imputation of missing values. The proposed approach outperforms alternative methods significantly for missing ratios >0.05 based on experimental results using real industrial datasets.

There are many data imputation strategies, from simple statistical methods such as mean imputation and regression imputation to more complex methods such as hot-deck imputation, which imputes the missing data by realistic scores that preserve the variable distribution[21]. Some widely-used imputation methods include: imputing using zero, mean, median, or mode; imputing using randomly selected value; and imputing using a model[22]. These techniques often impute a single and constant value for each variable without capturing or reflecting the relationship among the variables. This will likely result in an incorrect process-properties relationship.

Model-based imputation methods can be categorized into two types: those that make predictions for the missing values based on similar data points, and those that attempt to construct a global model to infer the missing data. The former includes algorithms such as k-nearest neighbors (kNN), while the latter encompasses deep learning neural networks.

The present study is focused on the investigation of the effect of different model-based imputation techniques on the process-structure relationship of the SLM Ti6Al4V dataset. The results of the imputation were evaluated to determine the best strategy for the dataset. This article will first present the methodology, followed by results and discussion about the different imputation methods, and finally the investigation of the imputed dataset.

2. Methodology

2.1. Imputation methods

2.1.1. k-Nearest neighbors (kNN) imputation

kNN imputation is one of the most common methods to impute missing values. It is used for both classification and regression problems[23]. The algorithm identifies k number of neighboring points using a distance metric and estimates the missing values using the values of these k neighboring observations[24].

The distance metric is generally Euclidean, and the function can be defined as

$$E(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad \text{(I)}$$

Where x_i and y_i are the point of interest and a case point from the dataset, and m is the number of input variables[25]. The process flow for the imputation is shown in Figure 2.

Since the kNN algorithm is non-parametric[23], there is no underlying assumption on the distribution of data, and hence, kNN is suitable for datasets with varied distributions.

Imputation was done using Scikit-learn's KNN Imputer class[26]. For calculation of the distance involving missing values, the coordinates of the missing value are ignored and the weights of the remaining coordinates

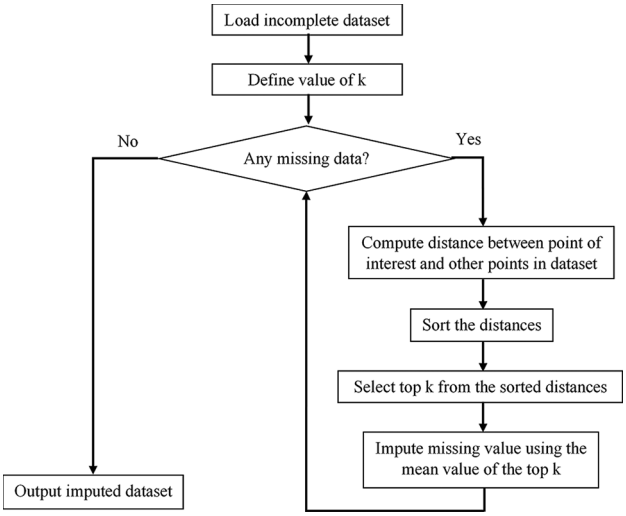

Figure 2. Process flow for k-nearest neighbor imputation.

are scaled-up[27]. The distance metric used to calculate the similarity between samples is the Euclidean distance. When calculating the distance involving missing values, the coordinates of the missing values are ignored. This means that when calculating the distance between two samples, only the coordinates where both samples have values are considered. The missing values are effectively treated as if they do not exist.

To account for the missing values, the weights of the remaining coordinates are scaled-up. This means that the distances between samples are adjusted to account for the missing values, so that samples that are similar in the remaining coordinates but have missing values in different locations are still considered similar. The scaling-up of weights is done by multiplying the weights of the remaining coordinates by a factor that is proportional to the number of non-missing coordinates in the samples being compared. Specifically, for each sample being compared, the weights of the remaining coordinates are divided by the proportion of non-missing coordinates in that sample. This means that the weights of the remaining coordinates are scaled-up by a factor equal to the reciprocal of the proportion of non-missing coordinates in the sample. This adjustment ensures that the distance metric takes into account the missing values in a meaningful way, without allowing the missing values to dominate the calculation. Each sample's missing values are imputed using the mean value from n_neighbors nearest neighbors, with n_neighbors = 5.

2.1.2. Multivariate imputation by chained equations

Multivariate imputation by chained equations (MICE)[28] is an imputation technique that iteratively imputes missing data for one variable modeled as a function of the other variables in a sequential fashion such that prior imputed values are used as part of the model in predicting subsequent variables. Hence, each variable can be modeled conforming to its distribution with continuous variables modeled using linear regression, while binary variables are modeled with logistic regression.

To carry out MICE, multiple copies of the dataset have to be created first. The following steps are then carried out on each copy of the dataset[29]:

(i) Missing values for each variable are imputed using non-missing values from the variable as a placeholder.
(ii) Set the imputed placeholders for one variable back to missing and model the selected variable as a function of the other variables. For each variable with missing values, the IterativeImputer class sets the imputed values for that variable to missing and models the selected variable as a function of the other variables using ExtraTreesRegressor. The model is trained on the complete cases, which are the cases where all variables are observed.
(iii) Using the fitted "ExtraTreesRegressor" model, predict, and impute missing values for the selected variable.
(iv) Repeat steps (ii) and (iii) for each variable in the dataset.
(v) The imputation cycle is repeated for 10 cycles, with the imputed values being updated at the end of each cycle.

The imputed copies of each dataset are then analyzed and the results combined using rules specific to the results[28], calculated using Rubin's Rules[30]. Rubin's Rule states that the estimated variance of the combined estimate is equal to the average of the within-imputation variance (the variability of the estimate within each imputed dataset) and the between-imputation variance (the variability of the estimates across the imputed datasets). To calculate the combined estimate, the point estimates from each imputed dataset are averaged, and the variance is calculated using Rubin's Rule. This approach accounts for the uncertainty due to missing data and provides estimates that are more accurate than those from the traditional complete case analysis.

Imputation was executed using Sckit-learn's IterativeImputer class, with the process flow as shown in Figure 3. Its implementation is similar to the R MICE package[28] but returns only one imputed dataset instead of multiple imputed datasets[31]. The estimator used for the sequential imputation was ExtraTreesRegressor, which builds an ensemble of regression trees, with default hyperparameters. Using ExtraTreesRegressor as the estimator for the IterativeImputer class, non-linear relationships between the variables in the dataset can be captured, which can result in improved imputations.

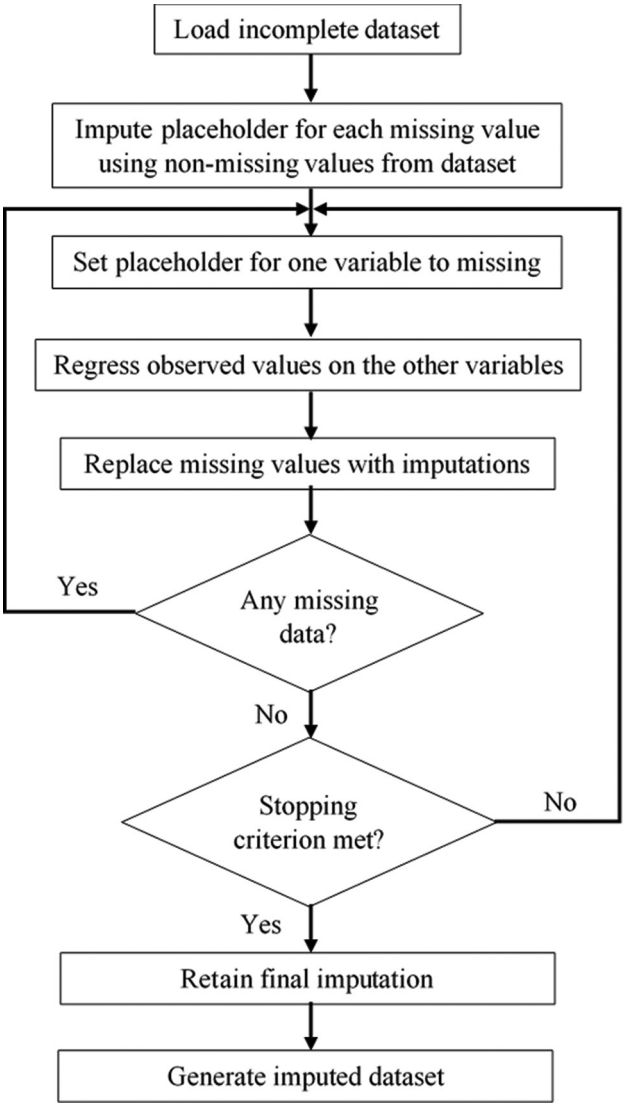

Figure 3. Process flow for graph imputation neural network imputation using IterativeImputer.

Since the Extra Trees algorithm is similar to Random Forest, which does not make any assumptions on the data[32], the distribution of the dataset to be imputed would not have a major impact on the imputation result.

2.1.3. Graph imputation neural network (GINN) imputation

GINN is a deep learning library that imputes missing data by first building a graph of similarities between complete values and then running an autoencoder with graph convolutions on top of that, with the schematics shown in Figure 4[33].

To develop the graph of similarities, each node of the graph is represented by a feature vector in the dataset, with Euclidean distance used to find the similarity between non-missing values[33]. Pruning is carried out twice by dropping all the connections below the 97.72th percentile, which was found to be a good value from experimental tests[33], for each row, leaving behind the most relevant nodes.

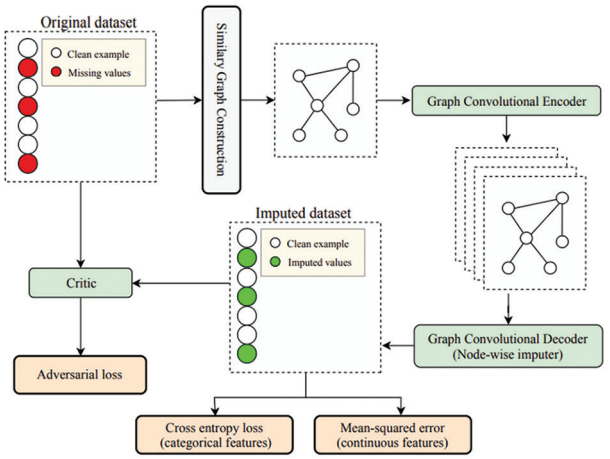

Figure 4. Schematics of graph imputation neural network imputation[27].

The graph convolution autoencoder used is defined as:

$$H=\mathrm{ReLU}\left(LX\Theta_1\right) \quad \text{(II)}$$

$$\hat{X}=\mathrm{Sigmoid}\left(LH\Theta_2\right) \quad \text{(III)}$$

Equation II is the encoder that takes the input X, with L as the Laplacian matrix related to the graph, and Θ_1 as a matrix of adjustable coefficients, and produces an encoded value H. Equation III is the decoder that provides a reconstructed imputed dataset $\hat{X}$ from the encoded H.

The graph convolution network is trained to minimize the loss function arising from errors in reconstruction as defined below:

$$L_A=\alpha\mathrm{MSE}\left(X,\hat{X}\right)+\left(1-\alpha\right)\mathrm{CE}(X,\hat{X}) \quad \text{(IV)}$$

Where MSE is the mean squared error for numerical variables, CE is the cross-entropy for categorical variables, and is an additional hyperparameter set as the ratio between numerical and categorical variables[33].

Adversarial training of the autoencoder is used in which a feedforward network is used as the critic to differentiate between imputed and non-imputed data, introducing an additional term to the loss function to include adversarial loss[33]. A more in-depth explanation of the GINN framework can be found in a previous work[33].

2.2. Cleaning of the dataset

Since there is a significant proportion of missing data in the dataset, initial data cleaning has to be carried out

before imputation as the information is of limited use if there is insufficient data.

The proportion of missing data is first calculated for each variable, and variables with more than 92% missing data are dropped (Table 1).

In general, process parameter variables have fewer missing values as the print parameters are normally reported regardless of the type of mechanical tests being conducted, whereas material property variables have high number of missing values as not every study has reported the same material properties. The threshold number of 92% is determined having considered the importance of the variables and the pattern of missing data. It is understood that the accuracy and reliability of the imputations may be lower when a large proportion of the data is missing. In general, imputation methods tend to perform better when the amount of missing data is lower and may struggle to accurately impute large amounts of missing data. Therefore, in our case, we attempted alternative method such as multiple imputation followed by a median approach to improve the accuracy. Of the remaining variables, scanning strategy and microstructure are dropped as they are too varied and unable to be generalized. Duplicate rows in the remaining dataset are then dropped. 401 datapoints were retained.

There are 18 variables retained: energy density (J/mm^3), exposure duration (μs), hatch spacing (μm), laser focus (mm), laser power (W), laser spot (μm), laser type (0 for continuous wavelength [cw], 1 for pulsed wavelength [pw]), layer thickness (μm), point distance (μm), scan speed (mm/s), density (%), elongation (%), microhardness (HV), macrohardness (HV), ultimate tensile strength (MPa), yield strength (MPa), Young's modulus (GPa), and porosity (%). Only two variables, laser power and laser type, do not have any missing values.

Cells that have a range of data inputted as a string (e.g., "60 – 180") are replaced with the mean values. As the exact value was not given for the variable, the use of mean values for the range is the only option, although it will lead to some degree of uncertainty. Cells with standard deviations (e.g., "0.12 ± 0.03") are replaced to retain only the numeric values in front.

2.3. Visualizing relationships in imputed dataset

After obtaining the imputed dataset using the median values obtained from the 3 algorithms, the process-properties linkages for SLM Ti64 can be obtained using data-mining through a self-organizing map (SOM). A SOM is an unsupervised machine learning model developed by Kohonen that reduces the dimensionality of an input space while maintaining its underlying structure[34]. This is especially useful to visualize large quantities of high dimensional data and model the relationship between them in a low, two-dimensional map, helping to advance the understanding of process-property relationships for materials.

The implementation of the SOM is from a Python package Tfprop_sompy, developed by Kikugawa and Nishimura, based on an open-source package SOMPY[35]. The training data were normalized by $x_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$ where x_{ij} is the i-th row of the j-th variable in the data, and μ_j and σ_j are the mean and standard deviations of the j-th variable, respectively. The size of the map was set to be 50 × 50 with the weights initialized using principal component analysis. Different sizes of the SOM were attempted and a map size of 100 × 100 was chosen such that each node of the map corresponds to at most one point of data in the dataset[36].

3. Results and discussion

3.1. Validation of the imputation models

Validation of the imputed values can be done through graphical plots that show the distribution of data, as well as numerical displays such as summary statistics of the imputed dataset[37].

Graphical evaluation of the imputed datasets is performed through data visualization using three plots: Boxplot, kernel density plot with histogram, and cumulative distribution plot. By comparing the statistical visualization plots, one can have an idea of the distribution of the imputed values and determine if they fall within expected boundaries.

Figure 5 shows the visualization plots for the energy density of the original dataset (observed) against the complete imputed dataset (imputed) for kNN imputations. The imputed values have close distributions to the original dataset and can be said to have reasonable values.

The three visualization plots for all the incomplete variables are plotted for each of the imputed datasets, and the imputed values for energy density, hatch spacing, laser spot, layer thickness, scan speed, elongation, microhardness, macrohardness, yield strength, and porosity for all three imputed datasets are found to be adequately close to the distribution of the original dataset (Figures S1–S3). However, the distributions for exposure duration, laser focus, point distance, and Young's modulus deviate from the original distribution to varying degrees for the different imputation techniques, with MICE (Figure 6) showing the greatest deviation followed by GINN (Figure 7) and then kNN (Figure 8).

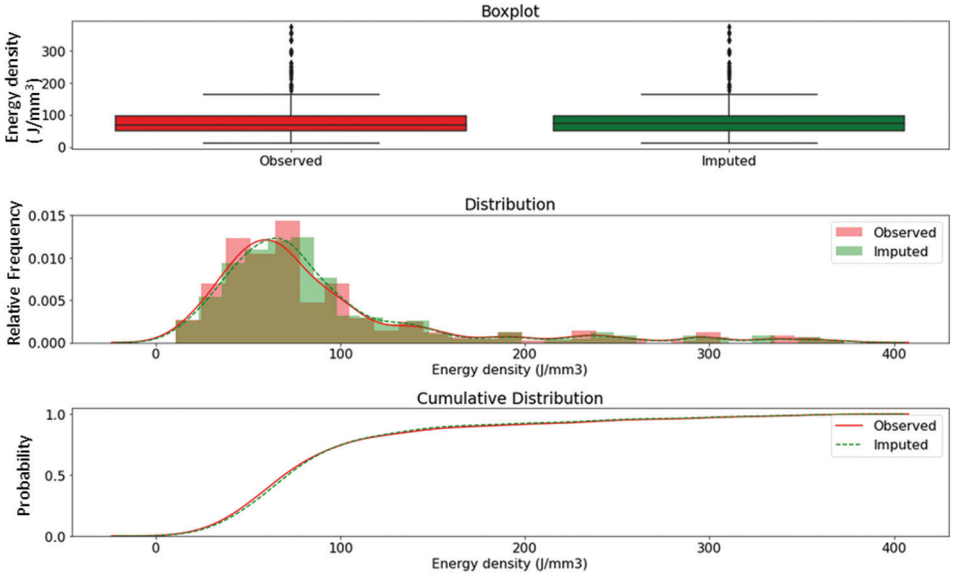

Figure 5. Graphs comparing the distributions of the observed (red) and imputed (green) energy density values. Top: Boxplot of the observed and imputed energy density values. Middle: Kernel density plot (line) with histogram (bars) of the observed and imputed energy density values. Bottom: Cumulative distribution plot of the observed and imputed energy density values.

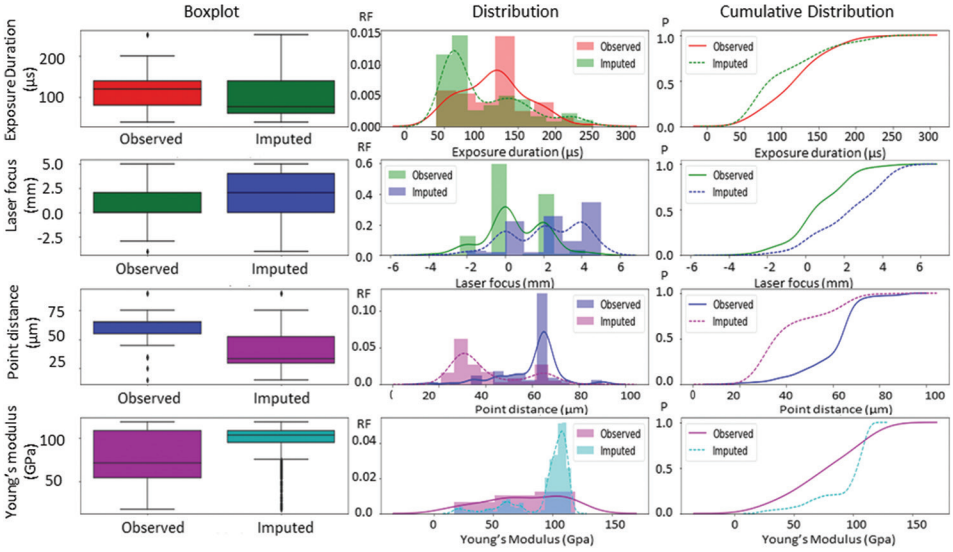

Figure 6. Graphs comparing the distributions of the observed and imputed values of exposure duration, laser focus, point distance, and Young's modulus for multivariate imputation by chained equations-imputed dataset, where RF is relative frequency and P is probability.

The variables exposure duration and point distance are only applicable to pulse wave laser for SLM. Since the dataset contains data from both pulse wave laser and continuous wave laser, the imputed values for these two variables would be for continuous wave laser parameter sets, which would not be relevant.

According to the literature[38], laser focus is a parameter that determines the laser spot size. However, the original literature that provided the values for laser focus did not state the laser spot size, and since there are several different ways to define the beam diameter[39], it is difficult to obtain the correct relationship between laser focus and laser spot for each observed value of laser focus. This could be a contributing factor, together with the limited data for laser focus, which led to the imputed values deviating from the original distribution.

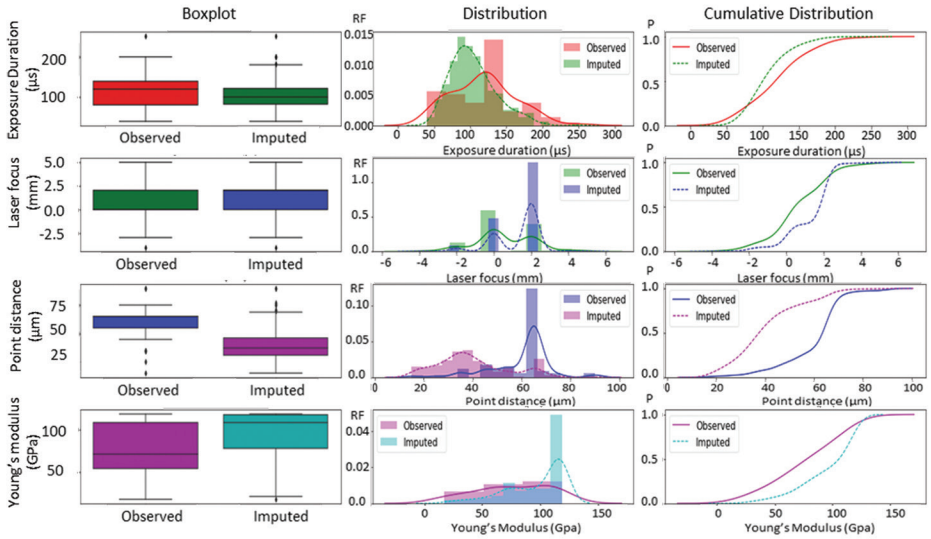

Figure 7. Graphs comparing the distributions of the observed and imputed values of exposure duration, laser focus, point distance, and Young's modulus for graph imputation neural network-imputed dataset, where RF is relative frequency and P is probability.

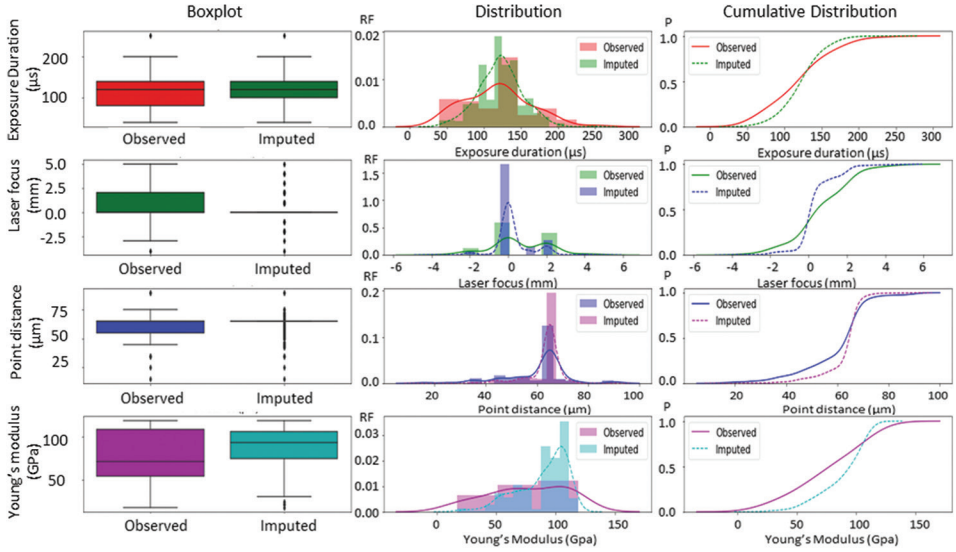

Figure 8. Graphs comparing the distributions of the observed and imputed values of exposure duration, laser focus, point distance, and Young's modulus for the k-nearest neighbor-imputed dataset, where RF is relative frequency and P is probability.

For the imputed values of Young's modulus, there appears to be a peak at around 120 GPa for all three imputed datasets. On further investigation of the imputed datasets, it was found that most of the imputed values close to 120 GPa correspond to sets of observed data that have similar values for the other parameters. Hence, it is not unreasonable for the range of values for the imputed Young's modulus values to be around 120 GPa. There is also limited observed data for Young's modulus (92% missing values), which could have contributed to the limited range of values imputed. To improve the distribution, more varied data with a larger range of Young's modulus values have to be obtained to allow the imputation algorithms to be more robust.

The complete visualization plots for all incomplete variables for all three imputed datasets can be found in the Supplementary File.

Occasionally, imputing more than 50% of the values may be required depending on the specific dataset and

research question. In this case, it is necessary to impute a large proportion of missing data to maintain a sufficient sample size and to include important variables in the analysis. However, imputing a high proportion of missing data can also increase the risk of bias and lead to inaccurate results. Therefore, it is important to carefully evaluate the validity of imputed data through various methods such as statistical summaries and comparison with observed data.

Statistical summaries can be used to validate imputed values, and Table 2 shows the compiled observed and imputed datasets. In the kNN-imputed dataset, the minimum and maximum values for all imputed variables remained unchanged from the original values. The mean and standard deviation of observed and imputed energy density values were similar (89.20 vs. 89.07 J/mm^3, and 68.05 vs. 65.14 J/mm^3, respectively). However, variables such as laser spot showed disparities in mean (125.56 vs. 106.03 µm) and standard deviation (133.86 vs. 97.57 µm), possibly due to differences in the proportion of missing data for each variable, with energy density having 364 observed values out of a total of 401, compared to only 194 observed values for laser spot.

Similarly, for the MICE-imputed and GINN-imputed datasets, the minimum and maximum values for all imputed variables did not change. There were also disparities in mean and standard deviations for variables laser focus and laser spot, possibly due to a large proportion of missing data, as discussed above. For the MICE-imputed dataset, the standard deviations of ultimate tensile strength and yield strength do differ (147.47 vs. 218.33 GPa, and 189.93 vs. 269.83 GPa, respectively) but given the proportion of missing data for these two variables, the imputed values may be reasonable. Thus, individual values of the imputed dataset have to be checked to ascertain if the imputations are sensible.

Table 1. Percentage of missing values for each variable

Variables	Missingness (%)
Laser power (W)	0.00
Laser type (0 for cw, 1 for pw)	0.00
Layer thickness (µm)	0.25
Hatch spacing (µm)	5.93
Energy density (J/mm^3)	9.14
Scan speed (mm/s)	14.81
Density (%)	43.46
Laser spot (µm)	51.60
Porosity (%)	69.14
Laser focus (mm)	73.09
EL (%)	79.51
Ultimate tensile strength (MPa)	79.51
Exposure duration (µs)	80.49
Point distance (µm)	81.98
Yield strength (MPa)	82.22
Macrohardness (HV)	88.15
Microhardness (HV)	90.12
Young's modulus (GPa)	92.35

Other than using the graphical and statistical methods to evaluate the imputed datasets, imputed values are also manually checked for any illogical values for the material properties: density and porosity values should add up to 100%, and the microhardness should be higher than the macrohardness[40]. Imputed values for process parameters should also fall within the processing window.

3.2. Comparison of imputation models

Comparing the distribution graphs, all three imputed datasets have relatively close distributions to the original dataset for the process parameters, as well as density and porosity variables. The discriminating features are the remaining variables, namely, elongation, microhardness, microhardness, ultimate tensile strength, yield strength, and Young's modulus. The model performs better for the processing parameters as they are deterministic and depend on fewer external factors. In addition, more datapoints are available for the processing parameters as they are reported in most of the studies. The material properties have a higher deviation because they have fewer datapoints as not every study focused on every aspect of material properties. There are also other factors such as different scan strategies, microstructures, and mechanical test conditions that are not captured in the dataset, leading to poorer imputation accuracy. As seen from the cumulative distribution plots (Figure 9) and distribution plots (Figure 10) of the three imputed datasets, GINN imputation results in the closest distribution to the original dataset.

The distribution of the kNN-imputed dataset has an acceptable deviation from the original distribution. However, an examination of the imputed dataset found that many imputed values for material properties are identical, even with different process parameters. The kNN algorithm did not manage to adequately capture the relationship between process parameters and material properties. Even so, it did successfully model the relationship between density and porosity, with all imputed values for these two variables adding up to 100%. There were also only a few instances where microhardness was lower than macrohardness.

Mean square error of the distributions is calculated and tabulated in Table 3. It was found that kNN performed

Table 2. Statistical summaries of observed and imputed datasets for kNN, MICE, and GINN for variables with missing values

Features	Observed					kNN imputation					MICE imputation					GINN imputation				
	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max
Energy density (J/mm^3)	364	89.20	68.05	11	373	401	89.07	65.14	11	373	401	92.93	67.56	11	373	401	85.32	67.23	11	373
Exposure duration (μs)	78	117.73	43.88	40	251	401	122.09	29.34	40	251	401	102.90	50.74	40	251	401	104.18	31.88	40	251
Hatch spacing (μm)	377	106.42	56.43	35	460	401	114.32	73.92	35	460	401	115.82	74.70	35	460	401	114.57	72.23	35	460
Laser focus (mm)	107	0.57	1.56	−4	5	401	0.28	0.96	−4	5	401	1.86	1.70	−4	5	401	1.30	1.22	-4	5
Laser spot (μm)	194	125.56	133.86	34	730	401	106.03	97.57	34	730	401	108.81	96.52	34	730	401	109.81	100.86	34	730
Layer thickness (μm)	400	45.30	34.89	20	200	401	45.27	34.85	20	200	401	45.32	34.85	20	200	401	45.25	34.86	20	200
Point distance (μm)	72	60.01	12.51	15	90	401	62.89	7.22	15	90	401	41.30	15.82	15	90	401	39.72	14.60	15	90
Scan speed (mm/s)	342	613.23	531.09	25	7000	401	638.93	514.52	25	7000	401	618.34	520.05	25	7000	401	580.02	501.83	25	7000
Density (%)	350	96.67	6.92	35	100	401	96.94	6.51	35	100	401	96.67	6.46	35	100	401	96.99	6.53	35	100
Elongation (%)	82	6.31	3.05	0	14.5	401	6.54	1.84	0	14.5	401	5.84	2.53	0	14.5	401	6.76	2.43	0	14.5
Microhardness (HV)	40	385.05	46.64	225	479	401	393.97	27.84	225	479	401	372.65	36.34	225	479	401	389.81	38.65	225	479
Macrohardness (HV)	47	366.31	26.84	262	403.5	401	372.27	12.76	262	403.5	401	356.86	28.90	262	403.5	401	358.81	34.32	262	403.5
Ultimate tensile strength (MPa)	82	1175.21	147.47	420	1407	401	1168.89	107.53	420	1407	401	1073.71	218.33	420	1407	401	1167.01	140.24	420	1407
Yield strength (MPa)	71	1049.56	189.93	0	1333	401	1040.13	125.32	0	1333	401	916.49	269.83	0	1333	401	1078.48	154.94	0	1333
Young's modulus (GPa)	31	75.64	32.28	17.12	119	401	89.08	20.84	17.12	119	401	95.03	22.62	17.12	119	401	96.59	25.66	17.12	119
Porosity (%)	350	3.33	6.92	0	65	401	3.06	6.51	0	65	401	3.34	6.46	0	65	401	3.02	6.53	0	65

N is the number of datapoints; maximum number of datapoints is 401 in this study; number of available datapoints for observed dataset is always smaller than 401 and may not be the same.
kNN: k-nearest neighbor, GINN: Graph imputation neural network, MICE: Multivariate imputation by chained equations

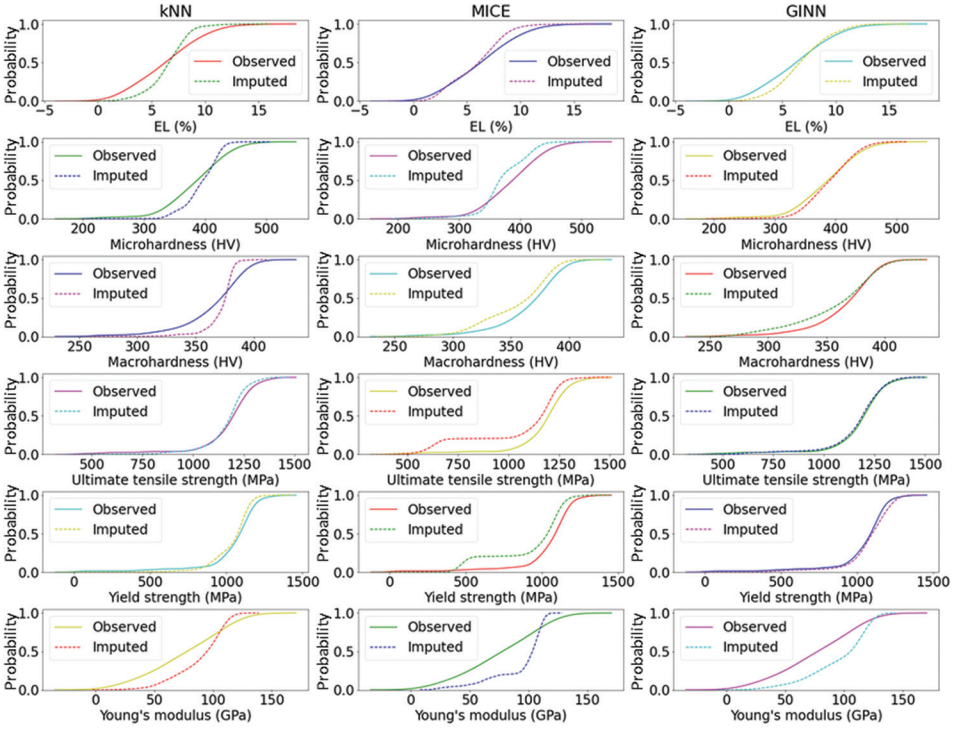

Figure 9. Cumulative distribution plots for k-nearest neighbor-imputed (left), multivariate imputation by chained equations-imputed (middle), and graph imputation neural network-imputed (right) datasets for selected incomplete variables.

the best for six variables, including exposure duration, hatch spacing, laser focus, layer thickness, point distance, and Young's modulus. GINN gave the best imputation for seven variables, including laser spot, density, elongation, microhardness, macrohardness, ultimate tensile strength, and yield strength. MICE only outperformed kNN and GINN in two variables, that is, scan speed and porosity. It is interesting to note that kNN performed better in imputing variables related to process parameters, whereas GINN performed better in variables related to material properties.

For the MICE-imputed dataset, the imputed values were much more varied, and the imputations showed a stronger relationship between the process parameters and material properties than the kNN-imputed dataset. While MICE performed better than kNN in that aspect, there were more instances where microhardness values were lower than macrohardness values (137 for MICE versus 73 for kNN). The sum of the imputed density and porosity values also do not add up to 100%, instead ranging from 99.92% to 100.12%.

The GINN algorithm can be said to have performed the best with the closest distribution to the original dataset and the most varied imputed values, although with 109 instances of lower values of microhardness compared to macrohardness, translating to 73% of the data showing the correct relationship. Most of the sums of imputed density and porosity values add up to 100% when rounded to 2 decimal places.

The performances of the models were validated by comparing the values obtained from the models with the actual values calculated from a known relationship. The laser scan speed is known to be related with other parameters by the equation below.

$$V = \frac{P}{Ehl} \tag{V}$$

where V is the laser scan speed, P is the laser power, E is the volumetric energy density, h is the hatch spacing, and l is the layer thickness. There are 48 datarows in the dataset that contains missing data for the laser scan speed while other parameters such as laser power, energy density, hatch spacing and layer thickness are available. These datarows can be used to validate the performance of the imputation models by comparing the imputed laser scan speed with the calculated laser scan speed using the known relationship as shown in Equation V. Figure 11 shows the performance of the models. It is found that the KNN model tends to overpredict the laser scan speed, while the GINN model tends to underestimate the laser scan speed. The MICE model is able to predict the laser scan speed correctly in

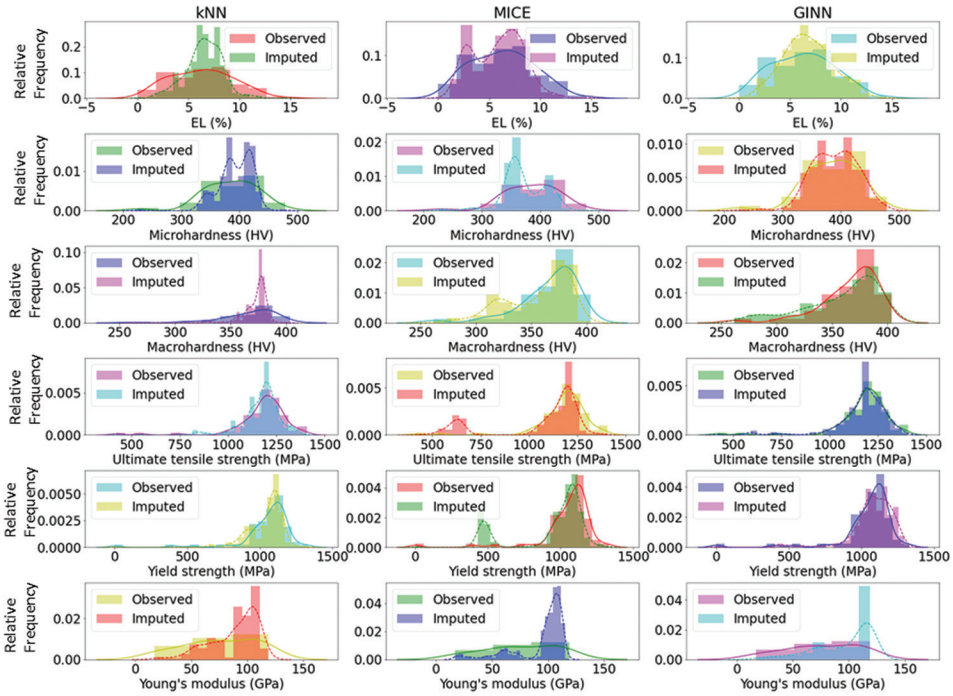

Figure 10. Distribution plots for k-nearest neighbor-imputed (left), multivariate imputation by chained equations-imputed (middle), and graph imputation neural network-imputed (right) datasets for selected incomplete variables.

Table 3. Comparison of the mean square error of the distribution for each variable

Features	kNN	MICE	GINN
Energy density (J/mm^3)	3.06E-07	1.71E-07[a]	3.11E-07
Exposure duration (μs)	1.64E-05[a]	4.46E-05	3.23E-05
Hatch spacing (μm)	1.32E-07[a]	7.72E-08	6.77E-08
Laser focus (mm)	2.90E-02[a]	3.03E-02	4.47E-02
Laser spot (μm)	1.19E-06	8.50E-07	2.97E-07[a]
Layer thickness (μm)	1.42E-09[a]	2.88E-09	4.26E-09
Point distance (μm)	1.23E-04[a]	8.56E-04	1.03E-03
Scan speed (mm/s)	1.11E-09	3.75E-10[a]	7.04E-10
Density (%)	1.77E-03	2.27E-03	1.82E-03[a]
Elongation (%)	3.87E-03	2.29E-03	1.81E-03[a]
Microhardness (HV)	1.50E-05	1.51E-05	4.81E-06[a]
Macrohardness (HV)	1.69E-04	2.99E-05	2.36E-05[a]
Ultimate tensile strength (MPa)	2.08E-07	4.52E-07	4.86E-08[a]
Yield strength (MPa)	9.12E-07	7.23E-07	2.30E-07[a]
Young's modulus (GPa)	2.24E-04[a]	2.26E-04	2.97E-04
Porosity (%)	6.01E-05	2.99E-05[a]	4.71E-05

[a]Lowest MSE among the three models. kNN: k-nearest neighbor, GINN: Graph imputation neural network, MICE: Multivariate imputation by chained equations

general, but the presence of outliers causes the relative mean square error (RMSE) (1.36) to be higher compared to the KNN (0.29) and GINN (0.16). GINN performs the best in terms of the RMSE.

To further validate the results, a composite parameter (energy density × hatch spacing) was used to evaluate the models. There are 19 datarows in the dataset that contains missing data for the energy density and hatch spacing, while other parameters of the energy density equation such as laser power, laser scan speed and layer thickness are available. Figure 12 shows the performance of the models. It is found that the MICE model tends to overpredict the composite parameter, while the GINN model tends to underestimate the composite parameter. The GINN model has a lowest RMSE (0.33) for the composite parameter, whereas MICE model has a higher RMSE (3.87) for the composite parameter.

To get the best out of the three models, further processing steps were taken. For each missing datapoint, the median of the imputed values obtained from the three models was taken as the final imputed value. This enhances the statistical confidence of the imputed value and reduces the chances of getting the outliers especially when dealing with small dataset. It resulted in a remarkable RMSE of 0.026, which is significantly lower than that of all three individual models (Figure 13). Nonetheless, it is believed that the models can be further improved with increased number of datapoints.

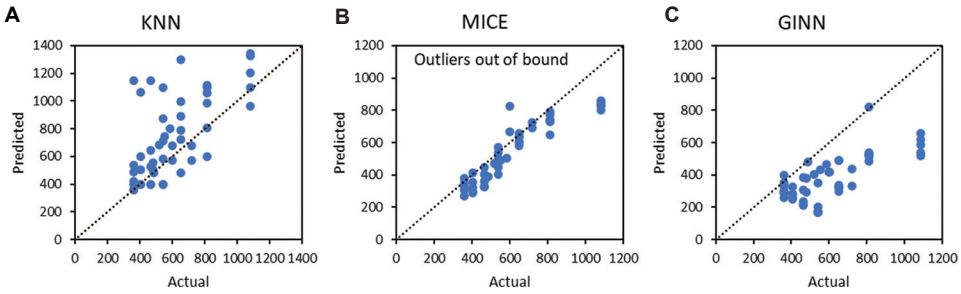

Figure 11. Comparison of the actual and predicted laser scan speed for various models. (A) k-nearest neighbor; (B) multivariate imputation by chained equations; (C) graph imputation neural network. See Supplementary File for the out-of-bound outliers.

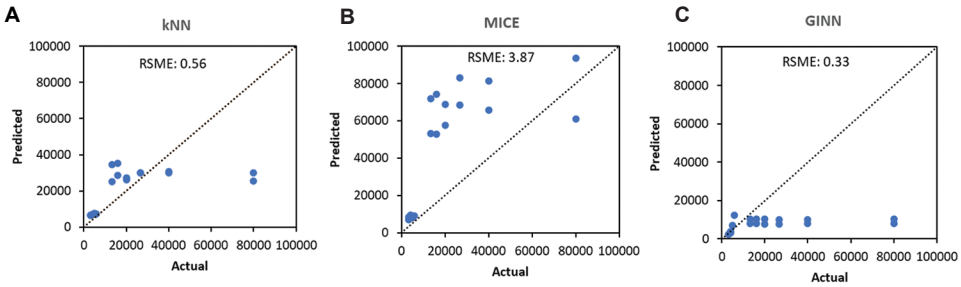

Figure 12. Comparison of the actual and predicted composite parameter (energy density × hatch spacing) for various models. (A) k-nearest neighbor; (B) multivariate imputation by chained equations; (C) graph imputation neural network.

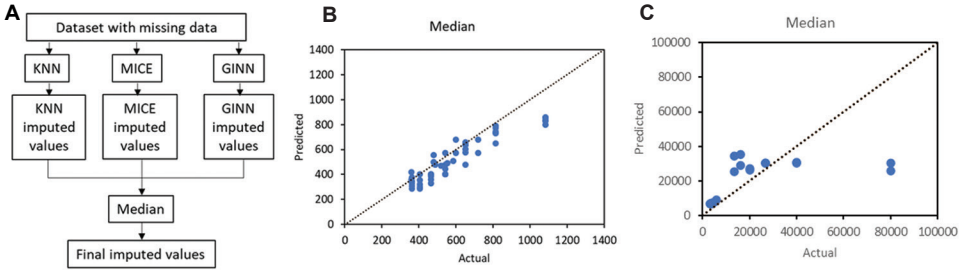

Figure 13. Strategy to improve quality of imputed dataset. (A) Median of the imputed values obtained from the k-nearest neighbor, multivariate imputation by chained equations, and graph imputation neural network. (B) The actual and predicted laser scan speed for the median of the imputed values. (C) The actual and composite parameter (energy density × hatch spacing) for the median of the imputed values.

The accuracy of the imputed results depends on several factors such as the quality of the original data, the imputation method used, and the amount and pattern of missing data. In this context, the imputation method used (GINN, MICE, and kNN) was found to be effective in reducing the overall RMSE and improving the accuracy of the imputed data. The final imputation strategy, which involved taking the median of the imputed values from the three models, further improved the accuracy of the imputed data.

While the results obtained in this study are promising, it is important to note that the suitability of the imputed data for wider implementation or industrial use depends on the specific context and requirements of the application. For example, if the imputed data are to be used for critical decision-making or safety-critical applications, a more rigorous validation process may be necessary to ensure the accuracy and reliability of the imputed data.

Data imputation is a crucial step in data analysis and modeling, especially when dealing with missing data. Imputation methods such as MICE and kNN can help to recover missing data and enable more robust and accurate data analysis. In addition, imputation can also help to reduce bias and increase the representativeness of the data, which can improve the quality of the insights and conclusions derived from the data.

The value of this work lies in its application of multiple imputation methods to a real-world dataset in the context

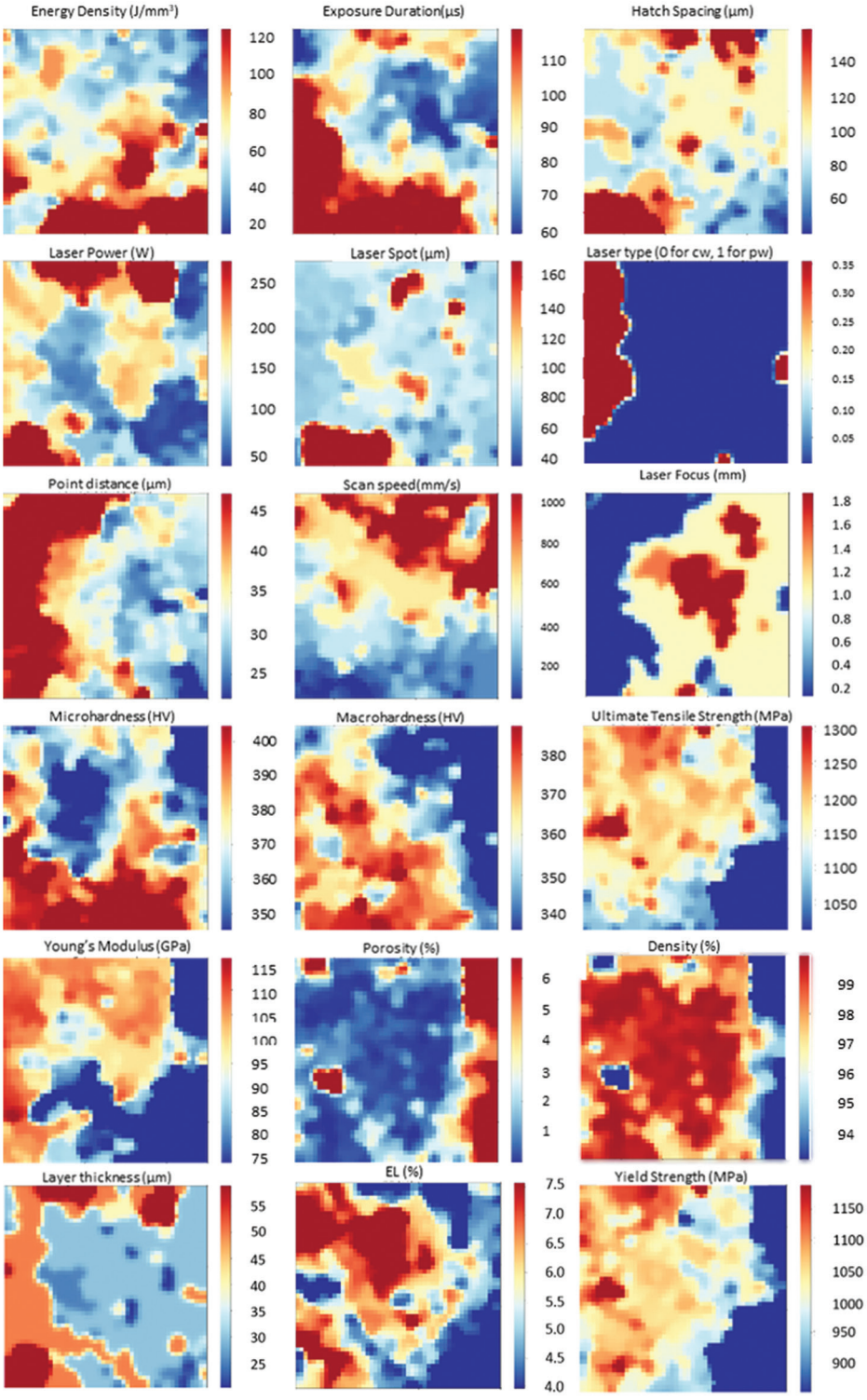

Figure 14. Heatmaps produced by self-organizing map for all 14 variables: 7 process parameters and 7 material properties. Red regions correspond to high values and blue regions correspond to low values.

of SLM. The study provides insights into the effectiveness of different imputation methods and highlights the importance of careful data cleaning and validation in ensuring the accuracy and reliability of the imputed

data. The study also demonstrates a practical approach to combining multiple imputation models to enhance the statistical confidence of the imputed data, which can be useful in other domains as well.

4. Visualizing relationships between variables in imputed dataset

Heatmaps for each variable was used to visualize the SOM trained, produced by linking each variable value to a node on the map grid (Figure 14). The relationships between the variables can be determined based on visual analysis, by comparing the locations of the red and blue regions that correspond to high and low values respectively. Based on the heatmaps, the following observations are made from the imputed dataset, some of which are well established relationship:

(i) Porosity is inversely related to Young's modulus and yield strength.
(ii) Scan speed is inversely related to microhardness and macrohardness.
(iii) Exposure duration is directly related to macrohardness.
(iv) Ultimate tensile strength, yield strength, and elongation are directly related.
(v) Energy density and scan speed are inversely related.
(vi) Laser power and Young's modulus are slightly directly related.

These relations provide insight into the process-property relationships in SLM Ti64 and can help users determine the process parameter window to obtain certain desired material properties. For example, to obtain a specimen with higher hardness, a lower scan speed, higher energy density and longer exposure duration should be used. A higher laser power is also likely to result in higher Young's modulus. The inverse relationship between the energy density and scan speed are found to be consistent with the energy density equation.

The inverse relationship between scan speed and microhardness in SLM Ti64 can be explained by several factors. First, high scan speeds can also lead to incomplete melting, resulting in the formation of unmelted or partially melted particles, which can act as a source of defects and lower the microhardness[41]. Second, there could be a change in microstructure of the printed Ti64 from a coarser equiaxed grains to a finer columnar grains as the scan speed increases. When the scanning speed is slow, the laser's slower movement increases both the energy input and stability of the molten pool. The elevated temperature in the molten pool creates adequate energy and nuclei for the epitaxial growth of columnar grains in the building direction. Wang *et al.* observed the coarsening of acicular structures in the samples produced at a speed of 250 mm/s compared to those produced at a speed of 1150 mm/s, and the shape was similar to that of heat-treated samples, as documented by Vilaro *et al.*[11,42]. As the scan speed increased to 1150 mm/s, the quantity of β-phase nanoparticles reduced, and only few white particles were dispersed across the fine acicular α (α′) grain boundaries.

There are several factors that can contribute to the relationship between laser power and Young's modulus in SLM of metals. One possible explanation is related to the changes in microstructure and grain size that can occur as a result of varying laser power levels. High laser power can lead to rapid melting and solidification of the material, resulting in smaller grain sizes and higher dislocation densities, which can contribute to an increase in Young's modulus.

It is noted from the heatmap that there is a slight correlation between laser power and Young's modulus. However, the effect of laser power on the Young's modulus in SLM Ti64 is not straightforward and depends on several factors. At low laser powers, the material experiences less thermal input and solidifies with a finer microstructure, resulting in a higher Young's modulus due to the increased strength of the material. However, as the laser power increases, the material is heated to a higher temperature, resulting in coarser microstructures due to increased grain growth and leading to a decrease in the Young's modulus. Furthermore, excessive laser power can result in porosity and defects in the material, which can significantly reduce the Young's modulus[10]. Therefore, the laser power should be optimized to achieve the desired microstructure and avoid porosity formation to ensure that the built parts have the required Young's modulus for their intended application.

Overall, it is important to consider the complex interplay between multiple process parameters and material characteristics that can affect mechanical properties in SLM. This also highlights the importance of including scan strategies and microstructures in the dataset for better generalization of the process-structure-properties relationship of SLM Ti64.

5. Conclusions

In this study, three model-based imputation techniques, kNN, MICE, and GINN imputations, were used to impute missing values in the Ti6Al4V dataset, which contained various process parameters and material properties obtained from multiple sources available in the literature. The results of the imputations were evaluated using graphical checks and statistical summaries to compare the imputed data with the original distribution before imputation. Among the three techniques, GINN

imputation gave the closest distribution to the original dataset and was the most accurate method, achieving the lowest RMSE. A median approach was used by taking the median of the imputed values from the three models. It was found that the median approach further improved the imputation accuracy by achieving RMSE of 0.026.

Data mining of the imputed SLM Ti64 dataset using SOM identified correlations between the process parameters and material properties. These correlations can be utilized to help users identify suitable process parameters for specimens with certain desired properties. While the material properties of monotonic yield strength and elongation at break are important, there are many other properties that could be of interest in the field of additive manufacturing. For example, fatigue strength, fracture toughness, creep resistance, and corrosion resistance are all important material properties that could be explored. In addition, exploring the relationship between process parameters and microstructure, such as grain size, could also provide valuable insights for optimizing the manufacturing process. The presented approach can also be applied to other databases to obtain new knowledge from the database.

However, a major limitation of the imputation methods is that a large proportion of missing data would lead to more inaccurate data imputation, and more manual checking of individual imputed values to ensure validity would be required. Imputation is not always appropriate and may introduce bias or lead to incorrect conclusions if the missing data is non-random or missing not at random. There is no specific quantitative threshold or limit to how much imputation can be performed before the results become meaningless. The validity of imputed data depends on the accuracy of the imputation models, the quality of the original data, and the degree and pattern of missingness.

In summary, imputation using the median approach was found to be the most accurate method for imputing missing data in the Ti6Al4V dataset, and the data mining approach using SOM identified correlations between process parameters and material properties. However, the study also highlights the need for more standardized testing and reporting, and the limitations of data imputation methods when dealing with a large proportion of missing data.

Acknowledgments

None.

Funding

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Medium-Sized Centre funding scheme.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

Conceptualization: Guo Dong Goh
Data curation: Jia Li Janessa Thong, Jia Jun Seah
Formal analysis: Jia Li Janessa Thong, Jia Jun Seah
Funding acquisition: Wai Yee Yeong
Investigation: Sheng Huang, Jia Li Janessa Thong, Jia Jun Seah
Methodology: Xi Huang
Project administration: Wai Yee Yeong
Resources: Wai Yee Yeong
Supervision: Wai Yee Yeong
Validation: Guo Dong Goh
Writing – original draft: Jia Li Janessa Thong
Writing – review & editing: Guo Dong Goh

Data availability

Data are available on request.

References

1. Liu S, Shin YC, 2019, Additive manufacturing of Ti6Al4V alloy: A review. *Mater Des*, 164: 107552.

   https://doi.org/10.1016/j.matdes.2018.107552

2. Elsayed M, Ghazy M, Youssef Y, *et al.*, 2019, Optimization of SLM process parameters for Ti6Al4V medical implants. *Rapid Prototyp J*, 25: 433–447.

   https://doi.org/10.1108/rpj-05-2018-0112

3. Roudnicka M, Bigas J, Vojtech D, 2020, Tuning porosity and mechanical properties of Ti6Al4V alloy additively manufactured by SLM. In: Key Engineering Materials. Vol. 865. Trans Tech Publications, Switzerland, p1–5.

4. Popovich A, Sufiiarov V, Borisov E, *et al.*, 2015, Microstructure and mechanical properties of Ti-6Al-4V manufactured by SLM. In: Key Engineering Materials. Vol. 651. Trans Tech Publications, Switzerland, p677–682.

5. Thijs L, Verhaeghe F, Craeghs T, *et al.*, 2010, A study of the microstructural evolution during selective laser melting of Ti-6Al-4V. *Acta Mater*, 58: 3303–3312.

   https://doi.org/10.1016/j.actamat.2010.02.004

6. Kuo C, Su C, Chiang A, 2017, Parametric optimization of density and dimensions in three-dimensional printing of Ti-6Al-4V powders on titanium plates using selective laser melting. *Int J Precis Eng Manuf*, 18: 1609–1618.

   https://doi.org/10.1007/s12541-017-0190-5

7. Pal S, Lojen G, Kokol V, *et al.*, 2018, Evolution of metallurgical properties of Ti-6Al-4V alloy fabricated in different energy

densities in the Selective Laser Melting technique. *J Manuf Process*, 35: 538–546.

https://doi.org/10.1016/j.jmapro.2018.09.012

8. Gong H, Rafi K, Starr T, *et al.*, 2013, The Effects of Processing Parameters on Defect Regularity in Ti-6Al-4V Parts Fabricated by Selective Laser Melting and Electron Beam Melting. In: Conference 24th Annual International Solid Freeform Fabrication Symposium.

9. Kasperovich G, Haubrich J, Gussone J, *et al.*, 2016, Correlation between porosity and processing parameters in TiAl6V4 produced by selective laser melting. *Mater Des*, 105: 160–170.

https://doi.org/10.1016/j.matdes.2016.05.070

10. Ali H, Ma L, Ghadbeigi H, *et al.*, 2017, *In-situ* residual stress reduction, martensitic decomposition and mechanical properties enhancement through high temperature powder bed pre-heating of Selective Laser Melted Ti6Al4V. *Mater Sci Eng A*, 695: 211–220.

11. Vilaro T, Colin C, Bartout JD, 2011, As-fabricated and heat-treated microstructures of the Ti-6Al-4V alloy processed by selective laser melting. *Metall Mater Trans A*, 42: 3190–3199.

https://doi.org/10.1007/s11661-011-0731-y

12. Qiu C, Adkins NJ, Attallah MM, 2013, Microstructure and tensile properties of selectively laser-melted and of HIPed laser-melted Ti-6Al-4V. *Mater Sci Eng A*, 578: 230–239.

https://doi.org/10.1016/j.msea.2013.04.099

13. Xu Y, Zhang D, Guo Y, *et al.*, 2020, Microstructural tailoring of As-selective Laser melted Ti6Al4V alloy for high mechanical properties. *J Alloys Compd*, 816: 152536.

https://doi.org/10.1016/j.jallcom.2019.152536

14. Pal S, Gubeljak N, Hudak R, *et al.*, 2019, Tensile properties of selective laser melting products affected by building orientation and energy density. *Mater Sci Eng A*, 743: 637–647.

https://doi.org/10.1016/j.msea.2018.11.130

15. Sun J, Yang Y, Wang D, 2013, Parametric optimization of selective laser melting for forming Ti6Al4V samples by Taguchi method. *Opt Laser Technol*, 49: 118–124.

https://doi.org/10.1016/j.optlastec.2012.12.002

16. Bartolomeu F, Faria S, Pinto E, *et al.*, 2016, Predictive models for physical and mechanical properties of Ti6Al4V produced by Selective Laser Melting. *Mater Sci Eng A*, 663: 181–192.

https://doi.org/10.1016/j.msea.2016.03.113

17. Fotovvati B, Namdari N, Dehghanghadikolaei A, 2018, Fatigue performance of selective laser melted Ti6Al4V components: State of the art. *Mater Res Express*, 6: 012002.

https://doi.org/10.1088/2053-1591/aae10e

18. Goh GD, Sing SL, Yeong WY, 2020, A review on machine learning in 3D printing: Applications, potential, and challenges. *Artif Intell Rev*, 54: 63–94.

https://doi.org/10.1007/s10462-020-09876-9

19. Steiner S, Zeng Y, Young TM, *et al.*, 2016, A study of missing data imputation in predictive modeling of a wood-composite manufacturing process. *J Qual Technol*, 48: 284–296.

https://doi.org/10.1080/00224065.2016.11918167

20. Wang Y, Li K, Gan S, *et al.*, 2019, Missing data imputation with OLS-based autoencoder for intelligent manufacturing. *IEEE Trans Ind Appl*, 55: 7219–7229.

https://doi.org/10.1109/TIA.2019.2940585

21. Andridge RR, Little RJ, 2010, A review of hot deck imputation for survey non-response. *Int Stat Rev*, 78: 40–64.

https://doi.org/10.1111/j.1751-5823.2010.00103.x

22. Jadhav A, Pramod D, Ramanathan K, 2019, Comparison of performance of data imputation methods for numeric dataset. *Appl Artif Intell*, 33: 913–933.

https://doi.org/10.1080/08839514.2019.1637138

23. Altman NS, 1992, An introduction to Kernel and nearest-neighbor nonparametric regression. *Am Stat*, 46: 175–185.

https://doi.org/10.2307/2685209

24. Imandoust SB, Bolandraftar M, 2013, Application of K-nearest neighbor (KNN) approach for predicting economic events: Theoretical background. *Int J Eng Res Appl*, 3: 605–610.

25. Wilson DR, Martinez TR, 2000, Reduction techniques for instance-based learning algorithms. *Mach Learn*, 38: 257–286.

https://doi.org/10.1023/A:1007626913721

26. sklearn.impute.KNNImputer-scikit-learn 0.23.2 documentation. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html [Last accessed on 2020 Oct 05].

27. sklearn.metrics.pairwise.nan_euclidean_distances-scikit-learn 0.23.2 documentation. Available from: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.nan_euclidean_distances.html [Last accessed on 2020 Oct 05].

28. Van Buuren S, Groothuis-Oudshoorn K, 2010, Mice: Multivariate imputation by chained equations in R. *J Stat Softw*, 45: 1–67.

https://doi.org/10.18637/jss.v045.i03

29. Azur MJ, Stuart EA, Frangakis C, *et al.*, 2011, Multiple imputation by chained equations: What is it and how does it work? *Int J Methods Psychiatr Res*, 20: 40–49.

https://doi.org/10.1002/mpr.329

30. Rubin DB, 1987, Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics). John

Wiley and Sons Inc., New York.

https://doi.org/10.1002/9780470316696

31. 6.4. Imputation of Missing Values-scikit-learn 0.23.2 Documentation. Available from: https://scikit-learn.org/stable/modules/impute.html#multiple-vs-singleimputation [Last accessed on 2020 Oct 05].

32. Shah AD, Bartlett JW, Carpenter J, *et al.*, 2014, Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. *Am J Epidemiol*, 179: 764–774.

https://doi.org/10.1093/aje/kwt312

33. Spinelli I, Scardapane S, Uncini A, 2020, Missing data imputation with adversarially-trained graph convolutional networks. *Neural Netw*, 129: 249–260.

https://doi.org/10.1016/j.neunet.2020.06.005

34. Kohonen T, 1982, Self-organized formation of topologically correct feature maps. *Biol Cybern*, 43: 59–69.

https://doi.org/10.1007/BF00337288

35. Moosavi V, Packmann S, Vallés I, 2014, SOMPY: A Python Library for Self Organizing Map (SOM). Available from: https://www.github.com/sevamoo/sompy [Last accessed on 2020 Oct 05].

36. Qian J, Nguyen NP, Oya Y, *et al.*, 2019, Introducing self-organized maps (SOM) as a visualization tool for materials research and education. *Results Mater*, 4: 100020.

https://doi.org/10.1016/j.rinma.2019.100020

37. Nguyen CD, Carlin JB, Lee KJ, 2017, Model checking in multiple imputation: an overview and case study. *Emerging Themes Epidemiol*, 14: 8.

https://doi.org/10.1186/s12982-017-0062-6

38. Metelkova J, Kinds Y, Kempen K, *et al.*, 2018, On the influence of laser defocusing in Selective Laser melting of 316L. *Addit Manuf*, 23: 161–169.

https://doi.org/10.1016/j.addma.2018.08.006

39. Slobodzian GE. White Paper-apples to Apples: Which Camera Technologies Work Best for Beam Profiling Applications, Part 2: Baseline Methods and Mode Effects. Available from: https://www.ophiropt.com/laser--measurement/knowledge-center/article/8065 [Last accessed on 2020 Oct 12].

40. Kuruvilla M, Srivatsan TS, Petraroli M, *et al.*, 2008, An investigation of microstructure, hardness, tensile behaviour of a titanium alloy: Role of orientation. *Sadhana*, 33: 235–250.

https://doi.org/10.1007/s12046-008-0017-2

41. Jiang PF, Zhang CH, Zhang S, *et al.*, 2021, Additive manufacturing of novel ferritic stainless steel by selective laser melting: Role of laser scanning speed on the formability, microstructure and properties. *Opt Laser Technol*, 140: 107055.

https://doi.org/10.1016/j.optlastec.2021.107055

42. Wang Z, Xiao Z, Tse Y, *et al.*, 2019, Optimization of processing parameters and establishment of a relationship between microstructure and mechanical properties of SLM titanium alloy. *Opt Laser Technol*, 112: 159–167.

https://doi.org/10.1016/j.optlastec.2018.11.014