



Analysis of Data Generation and Preparation for Porosity Prediction in Cold Spray using Machine Learning

Martin Eberle^{1,2} · Samuel Pinches¹ · Max Osborne² · Kai Qin³ · Andrew Ang¹ 

Submitted: 6 October 2023 / in revised form: 2 March 2024 / Accepted: 4 March 2024
© The Author(s) 2024

Abstract Cold spray is an additive manufacturing and coating process in which powder particles are accelerated to supersonic speeds without melting them and then deposit on a surface to form a layer of a coating. Process parameters and materials affect the characteristics of manufactured parts and therefore must be chosen with care. Machine learning (ML) techniques have been specifically applied in additive manufacturing for tasks such as predicting and characterizing porosity. Machine learning algorithms can learn how a variation in the input spray parameters affects annotated output data, such as experimentally measured part properties. In this work, a dataset was developed from experiments reported in published

academic papers, to train ML algorithms for the porosity prediction of cold spray manufactured parts. Data cleaning steps, such as null value replacement and categorical feature handling, were applied to prepare the dataset for the training of different ML models. The dataset was split into training and testing portions, and floating feature selection and hyperparameter optimization were performed using parts of the training set. A final evaluation of all trained models, using the test portion of the dataset, showed that a prediction accuracy with an average deviation of 0–2% porosity of the predicted values compared to the true values can be achieved.

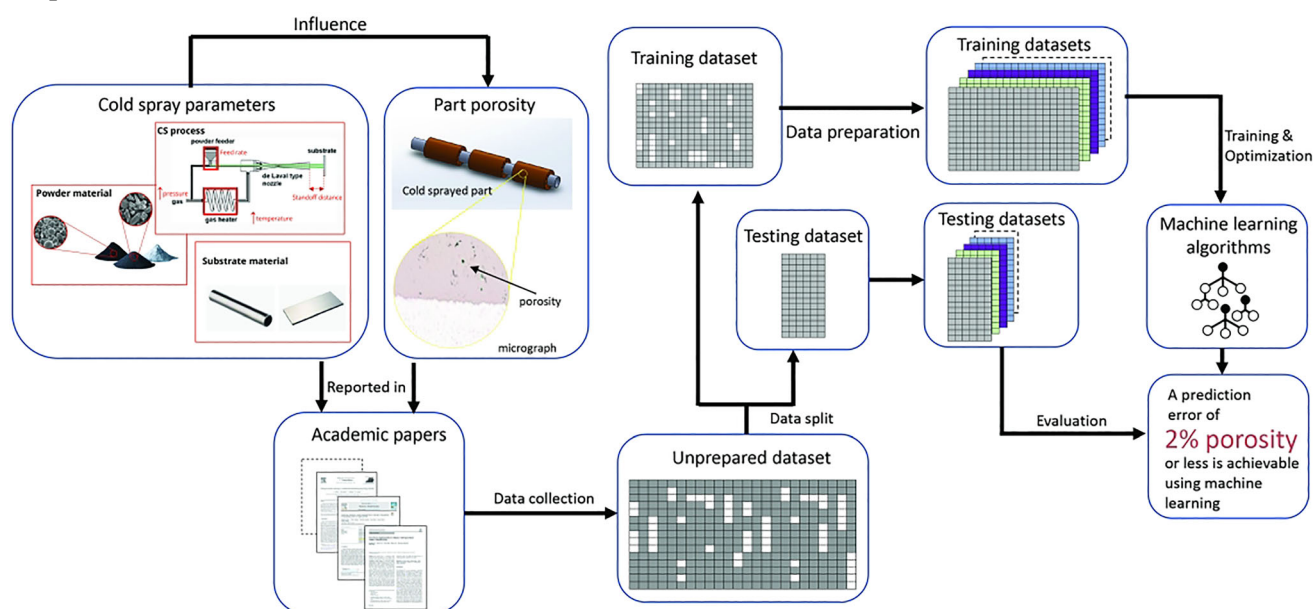
✉ Andrew Ang
aang@swin.edu.au

¹ ARC Training Centre on Surface Engineering for Advanced Materials (SEAM), School of Engineering, Swinburne University of Technology, Hawthorn, VIC 3122, Australia

² Titomic Limited, TKF Melbourne Bureau, 1/371 Ferntree Gully Road, Mount Waverly, VIC 3149, Australia

³ Department of Computing Technologies, Swinburne University of Technology, Hawthorn, VIC 3122, Australia

Graphical Abstract



Keywords additive manufacturing · cold spray · data-driven models · machine learning · porosity

Introduction

Cold spray (CS) is a solid-state additive manufacturing and coating process, in which powder particles are accelerated to supersonic speeds, after which they are deposited on a surface without melting. The underlying bonding mechanisms are local metallurgical bonding and mechanical bonding which are caused by localized plastic deformation between the particles and/or the substrate. The plastic deformation and therefore the formation of a cold spray deposit rely largely on the particle kinetic energy before impact (Ref 1).

A vast array of more than 30 process parameters are responsible for determining the characteristics of the manufactured parts, such as the temperature and pressure of the accelerating gas, spray nozzle geometry, distance of the spray nozzle to the sprayed surface, and ultimately, the motion path of the spray nozzle. Further, the characteristics and conditions of the spray powder material and the substrate material also affect the properties of the final manufactured parts (Ref 2, 3). Due to this complexity, the optimization of the CS process poses significant challenges.

The resulting level of porosity is a critical property of a cold-sprayed deposit as it is directly related to coating or deposit quality and performance. For some applications, it

can be desirable and intentional to achieve a high level of porosity, whereas in other cases a highly dense deposit is required. Figure 1 shows two cross-sectional microscope images of CP titanium samples with a low porosity of 0.7%, achieved with a gas pressure of 49 bar and a gas temperature of 1000 °C (a) and a higher porosity of 5.0% achieved with a gas pressure of 50 bar and a gas temperature of 880 °C (b).

An example of an application for porous cold-sprayed deposits is orthopedic implants made of titanium. A porous structure permits bone tissue ingrowth and provides a reliable fixation at the bone implant interface. Sun et al. (Ref 4) were able to cold spray a titanium coating with a porosity of 48.6% on a titanium substrate suitable for the aforementioned application.

For other applications a porous coating microstructure is detrimental. For example, when applying a protective surface coating, a porous coating may inadequately protect the underlying material if it is exposed to a corrosive environment, leading to corrosion damage of the structure (Ref 5). Further, the mechanical properties of metal parts are adversely affected by porosity. Therefore, both the total porosity and pore size of load bearing components must both be kept to a minimum (Ref 6).

The porosity of a cold-sprayed coating or deposit is highly sensitive to both material characteristics and processing parameters (Ref 7). Table 1 provides examples of relevant cold spray process parameters and their reported effects on the resulting porosity. It should be noted that the following identified relationships are specific to their

Fig. 1 Cross-sectional microscope images of CP titanium samples with low porosity (a) and high porosity (b), manufactured at Titomic Ltd. (Melbourne, Australia)

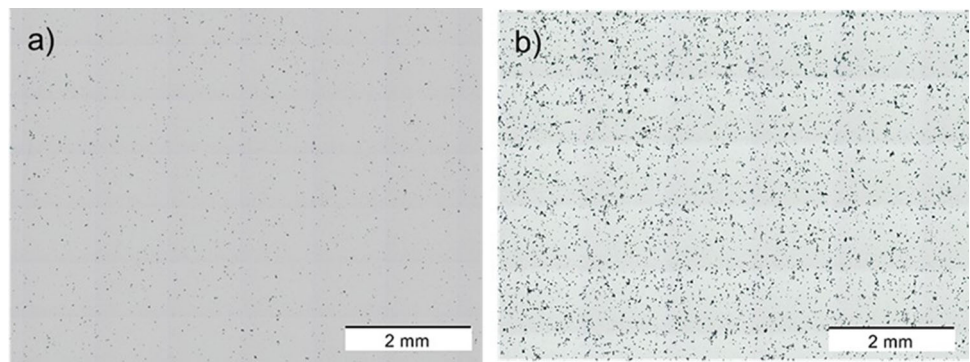


Table 1 Parameters affecting the porosity in cold spray

Parameter	Parameter	Effect on porosity	Type of study	Powder material	Morphology	Reference
Nozzle traverse speed	Increase	Increase	Experimental	CP titanium	Spherical	Ref 8
Nozzle traverse speed	Increase	Increase	Experimental	CP titanium	Irregular	Ref 9
Coating/deposit thickness	Increase	Increase	Experimental	Ti-6Al-4V	Spherical	Ref 10
Particle impact velocity	Increase	Decrease	Experimental	CP titanium	Spherical	Ref 11
Particle impact velocity	Increase	Decrease	Experimental	CP titanium	Irregular	Ref 12
Particle impact velocity	Increase	Decrease	Simulation	Ti-6Al-4V	Spherical	Ref 13
Particle impact temperature	Increase	Decrease	Experimental	CP titanium	Spherical	Ref 11
Particle impact temperature	Increase	Decrease	Experimental	CP titanium	Irregular	Ref 12
Particle impact temperature	Increase	Decrease	Simulation	Ti-6Al-4V	Spherical	Ref 13
Powder feed rate	Increase	Increase	Experimental	CP titanium	Irregular	Ref 14
Gas molecular weight	Increase	Increase	Experimental	CP titanium	Irregular	Ref 15
Gas pressure	Increase	Decrease	Mathematical	CP titanium	No info	Ref 16
Gas temperature	Increase	Decrease	Mathematical	CP titanium	No info	Ref 16
Gas temperature	Increase	Decrease	Experimental	304 stainless steel	Irregular	Ref 17
Particle size	Increase	Increase	Experimental	CP titanium	Spherical	Ref 11
Particle size	Increase	Increase	Experimental	Copper	Spherical	Ref 18
Particle size	Increase	Increase	Experimental	CP titanium	Irregular	Ref 15
Particle size	Increase	Decrease	Experimental	CP titanium	Irregular	Ref 19
Particle size	Increase	Increase	Experimental	316 stainless steel	Spherical	Ref 20
Spray angle*	Increase	Increase	Experimental	CP titanium	Spherical	Ref 21
Spray angle*	Increase	Decrease	Experimental	CP titanium	Irregular	Ref 9
Nozzle standoff distance	Increase	Increase	Experimental	CP titanium	Irregular	Ref 15
Nozzle standoff distance	Increase	Increase	Mathematical	CP titanium	No info	Ref 16
Substrate material	Change	No effect	Experimental	CP titanium	Irregular	Ref 12
Substrate temperature	Increase	No effect	Simulation	Ti-6Al-4V	Spherical	Ref 13
Substrate temperature	Increase	Decrease	Experimental	CP titanium	Spherical	Ref 8
Substrate roughness	Increase	No effect	Experimental	Al-Mg	Irregular	Ref 22
Particle surface reactivity	Increase	Increase	Experimental	CP titanium	Angular	Ref 23

*Normal: 90°.

particular study conditions and parameters, and as such, these relationships may not have been verified to hold more generally.

While the changing of individual CS process parameters has observable effects, many of these parameters have

complex interdependencies. For instance, the reported contradictory trends in studies by Zahiri et al. (Ref 15) and Marrocco et al. (Ref 19), where one observes increased porosity with larger powder particle size while the other reports the opposite, can be attributed to variations in

experimental conditions. Factors such as different cold spray systems, accelerating gases, and the potentially non-monotonic nature of relationships between spray and material characteristics and porosity contribute to the observed discrepancies.

In another example, Magarò et al. (Ref 24) observed that the effect of the nozzle traverse speed on the mechanical properties of stellite-6 coatings is more prominent when low values of gas temperature and pressure are imposed, whereas for the highest ones smaller differences are noticeable.

These nuanced influences suggest that authors may indeed measure different trends under distinct experimental setups and that makes it challenging to optimally control or optimize the process, to achieve specific final process characteristics within manufactured parts or coatings.

Porosity in materials can be measured through various methods, including image analysis, x-ray Microtomography (XMT), and the Archimedes principle. Image analysis technique involves the binarization of microscope images of sample cross-sections with software like ImageJ and determining porosity as the ratio between black and white pixels. XMT employs x-ray imaging to capture detailed, cross-sectional views of materials, providing insights into their internal structures. The Archimedes principle involves measuring the displacement of fluid when a sample is submerged, offering a way to determine the volume and, consequently, the porosity of the material (Ref 12, 13, 25).

Computational models offer an alternative approach to determine the porosity without the need to manufacture physical samples. Terrone et al. (Ref 26), Song et al. (Ref 13), and Weiller and Delloro (Ref 3) employed coupled Eulerian–Lagrangian (CEL) frameworks to simulate multiple particle impact in CS and quantify the resulting porosity in the material. Terrone et al. (Ref 26) focused on multi-material titanium-copper and titanium-aluminum combinations and on the prediction of coating porosity after removal of a sacrificial aluminum or copper phase, varying volume fractions of these sacrificial materials. The porosity was assessed through a simulated chemical etching process, selectively removing sacrificial materials connected to the deposit surface and transforming those in direct contact with the external environment into pores.

Song et al. (Ref 13) analyzed Ti-6Al-4V powder particle impacts on Ti-6Al-4V substrates and found that the substrate temperature had a negligible impact on deposit porosity, which was mainly influenced by particle temperature and velocity. Porosity, represented by the average Eulerian volume fraction (EVF) void percentage, was determined by analyzing a cubic volume within the deposit and counting void elements.

Weiller and Delloro (Ref 25) simulated the impact of a combination of aluminum and alumina particles on an

aluminum alloy substrate. To assess the porosity level in the simulation results, a method based on the calculation of successive convex hulls was employed to avoid user-dependent sampling volume choice that might result in non-representative or too small volumes for the porosity evaluation in the deposit. Comparing the porosity predictions with x-ray microtomography (XMT) results revealed notable differences, which were suspected to be due to different volume fraction sizes analyzed in simulation and experiments. An increase in the fraction volumes generated in the simulation to match the sizes of those in the experiments was expected to improve the prediction accuracy, but it would lead to prohibitive computational costs.

The mentioned models share common limitations, as they are tailored to specific use cases, restricting their applicability to a broad range of spray scenarios. Further the predicted porosity accuracy is influenced by mesh size and analyzed volume fraction, with smaller mesh sizes and larger volume fractions achieving higher accuracy but increasing computational time, limiting the practical usability of the models (Ref 13, 25).

Machine learning (ML) models offer an alternative data-driven approach for predicting outcomes by learning from examples, capable of handling multi-dimensional and heterogeneous data to detect both linear and nonlinear relationships (Ref 27). Multidimensionality refers to the number of input and output parameters, and the heterogeneity refers to a high variability of data types of the parameters in the dataset (Ref 28). ML models can deliver fast prediction result within seconds [29, 30] and have been successfully implemented in CS to predict process outcomes (Ref 1-5).

Valente et al. (Ref 31) used a decision tree classifier ML algorithm to classify the flowability of powders used in CS from powder features such as area, perimeter, ellipse ratio, and circularity. Testing the algorithm with unseen data led to an accuracy of 77.31%. Their dataset only comprised 21 datapoints with 29 features.

Ikeuchi et al. (Ref 32) used a neural network (NN) model to predict the shape of single-track profiles for cold spraying titanium from the input variables spray angle, traverse speed, and standoff distance. The model's mediocre mean absolute percentage error (MAPE) of 8.342% on the test data was attributed to an insufficient data quantity. Therefore, in their following study they added virtual input–output subsets generated by a Gaussian function model to the measured profile shape data. This technique enabled them to improve the prediction accuracy compared to the previous model, and they achieved a MAPE of 1.23% on the test data set (Ref 33).

Liu et al. (Ref 34) utilized a NN to predict multi-layer profiles of cold-sprayed deposits on diverse substrate morphologies. Inputs included stand-off distance, gun

traverse speed, deposition cycle count, and 206 points from the previous layer's profile, while coating profile data served as targets. The NN model achieved a relative error (RE) within 10% when compared to experimental profiles. Discrepancies were attributed to assumptions in the model, not accounting for substrate abrasion and measurement inaccuracies of the 3D coating profiler used to generate the data (Ref 34).

Wang et al. (Ref 35) developed a NN model to predict critical velocity using 10 material features as model input. The 8 datapoints for training their models were taken from the literature. While the developed NN model demonstrated a high prediction accuracy of 96.45% when tested with the property data of two new materials, it is important to note that assessing the model's performance with a limited dataset of only two datapoints does not provide a comprehensive validation of its predictive capabilities.

The examples show that the prediction accuracy of the models is depending largely on the dataset size and quality of the data. Further these dataset characteristics determine the application range of the models.

This paper aims to develop for the first time a ML model that can be applied to predict the porosity in cold spray metal deposits using process parameters and spray powder and substrate properties as inputs. The dataset necessary to train ML algorithms for this task was developed from experiments reported in published academic papers. With this strategy, a larger pool of data can be generated, and a larger universal applicability of the models can be achieved, helping future users of the models to apply it for a wide range of CS process scenarios. While this approach offers the benefit of models with a wide application range, there is a potential drawback associated with using literature data. The utilization of such data introduces a higher level of uncertainty regarding its accuracy and quality, which in turn may negatively impact the performance of the model. For instance, as different authors reported different sets of parameters, in the resulting combined dataset there were numerous parameters with incomplete data available.

To exploit the advantage of a dataset generated from the literature and to overcome its drawbacks, several data preparation techniques were applied to pinpoint the most effective ways to prepare a high-quality dataset. In “[Machine Learning Algorithms](#)” section of this paper, the applied ML algorithms are described. “[Data Preparation](#)” section is about the extensive data preparation strategies that were applied, including a parameter selection strategy, missing value, and categorical feature handling strategies. In “[Method](#)” section, methods of measuring the feature importance, null value replacement accuracy, and prediction accuracy are introduced. Finally, in “[Results and](#)

[Discussion](#)” section the results of the applied methods are presented including a description of limitations and suggestions for future works.

Machine Learning Algorithms

Machine learning is an artificial intelligence (AI) technique that allows a computer to automatically learn from data and make decisions or predictions without being explicitly programmed [36]. Once an ML algorithm has been trained, it is presented with a test data set to determine how well it can identify the samples that it has never seen before [27].

In this work, the bagging ensemble method random forest (RF) and the boosting ensemble methods extreme gradient boosting (XGB) and categorical boosting (CatB) were deployed on a dataset to make predictions of the porosity in cold spray. A RF is an ensemble learning method where multiple decision trees are trained on random subsets of the training data with replacement. Each decision tree makes predictions independently, and the final prediction is determined by averaging the predictions of all trees, resulting in improved accuracy and robustness compared to individual decision trees (Ref 37).

XGB, introduced by Chen (Ref 38), is an advanced boosting ensemble method. Specifically, XGB uses a more regularized model formalization which can reduce the sensitivity of predictions to individual observations. Thereby, the risk of overfitting to the training data can be reduced, and the overall performance of the model can be improved. Further XGB focuses on the computational power, by parallelizing the tree formation leading to speed gains in the training of the model. Another advantage of the XGB algorithm is that it can handle missing values in the dataset and therefore does not necessitate the application of missing value replacement methods (Ref 38).

CatB is a boosting variant developed by researchers at Yandex (Ref 40) that stands out for its ability to handle categorical values. This is achieved through an ordered target encoding (OTE) technique. In OTE, each instance's category is replaced with the mean target value of instances that come before it in the dataset. By avoiding reliance on information from the entire dataset, the model is less likely to overfit the training data, enhancing its generalization to new, unseen data. CatB also applies an ordering principles to the development of the tree structure and the calculation of the residuals, known as ordered boosting, which leads to a further mitigation of the risk of overfitting (Ref 39).

The mathematical details of the RF, XGB and CatB algorithms can be found in Breiman (Ref 37), Chen (Ref 38), and Prokhorenkova (Ref 39), respectively.

Data Preparation

Here, the data for training and testing the ML models were generated from the information provided in cold spray academic papers about process parameters and the associated part properties. The inclusion of instances to the dataset was constrained to non-heat treated, as-sprayed samples that were produced with a one-component metal powder type. The information provided by each paper varied significantly leading to a non-negligible number of missing values in the data set. Due to the high number of missing values for some of the parameters, a minimum of 35% completeness was set as a threshold for inclusion to the dataset. The completeness of a parameter was derived as the ratio of non-null count of the respective parameter to the total number of experiments. Therefore, from 35 papers 242 porosity measurements and data about 14 relevant process parameters that were used to produce the respective samples were included in the dataset. The percentage of completeness for the entire dataset was 84%. Figure 2 illustrates the structure of the dataset. The vertical axis represents the parameters of the dataset (such as gas temperature, gas pressure, etc.), and the horizontal axis represents the experiments reported in an academic paper. Therefore, each field in this matrix represents a specific parameter value that was used in the respective experiment reported in the paper. In some cases, the information about certain parameters was not shared in the article, leading to a missing value in the data matrix. The difference between fields with information content and fields with missing data is indicated in Fig. 2. Black fields represent parameters with information content, white fields represent missing data. Table 2 gives an overview of the 14 parameters and the extent of missing data.

Parameter Selection

The success of a machine learning project is highly dependent on the selection of parameters. Considering the available dataset with $n = 14$ parameters and the option to choose from $k = \{1, \dots, 14\}$, there are a total of 16,383 possible combinations of parameters.

$$\text{Number of combinations} = \sum_{k=1}^{14} \frac{n!}{(n-k)! \cdot k!} \quad (\text{Eq 1})$$

where n : elements to choose from k : elements chosen.

To reduce the number of models to be tested, only selected parameter combinations were tested. In this case, a floating dataset selection was constructed, with a set of variables selected to correspond with a varying range of variable data completeness. Table 3 shows the resulting configurations (0-10) of the dataset.

The initial configuration 0 is a dataset including the parameters powder material, substrate shape, process gas type, and gas pressure. For each following configuration (1-10), one additional new parameter was added (see parameters from Table 2). The order in which the new parameters were added was depending on the non-null count of the respective parameter (see Table 2, columns 'non-null count'). Parameters with a high non-null count were added first. With this method, a range of dataset configurations was created. Starting with configurations without missing values (configurations 0, 1, 2), each following configuration (configurations 3-10) exhibits an increasing number of missing values.

Adding more parameters can contribute to the accuracy of the predictions of a machine learning algorithm if there is a strong correlation between the respective parameter and the target parameter (Ref 40).

On the other hand, in this case the inclusion of a parameter leads to an increase in the number of null values in the dataset and all gaps in the dataset had later to be replaced using a null-value replacement strategy. Depending on how close the null-value replacements are to the unknown true values, this can result in a lower accuracy of the predictions by the model (Ref 41).

Handling of Missing Values

The validity of the application of missing data handling strategies for an incomplete dataset requires meeting certain assumptions about the reason for missingness. These reasons are referred to as the missing data mechanisms (Ref 42). A value of a variable X_1 is said to be missing completely at random (MCAR) if the probability that X_1 is



Fig. 2 Structure of dataset developed from academic papers. Black fields represent available information, whereas white fields represent missing data

Table 2 Overview of the parameters of the dataset developed from experiments reported in academic papers

Parameter category	Parameter	Unit	Non-null count	Parameter type
Powder	Powder material	...	242	Categorical
	Powder morphology	...	207	Categorical
	Avg. particle size (volume mean)	µm	111	Numerical
Substrate	Substrate material	...	236	Categorical
	Substrate shape	...	242	Categorical
System	Cold spray system	...	149	Categorical
Nozzle properties	Nozzle type	...	97	Categorical
Nozzle/robot movement	Traverse speed	mm/s	179	Numerical
	Spray angle	°	242	Numerical
	Standoff distance	mm	231	Numerical
Powder feeder	Powder feeder rate	g/min	164	Numerical
Process gas	Process gas type	...	242	Categorical
	Gas pressure	MPa	242	Numerical
	Gas temperature	°C	242	Numerical

Avg. = average.

missing is not related to a specific value of X_1 or the observed values of other variables X_n (Ref 42). That means that the missing observations are just a random subset of all observations and there is no systematic difference between the missing and observed data. The missing observations of an explanatory variable X_1 are called missing at random (MAR) if the likelihood that the values of X_1 are missing depends on the observed values of another variables X_n , but not values of the variable X_1 (Ref 42). In other words, the missingness of a field can be explained by the values in other columns, but not from that column. Missing data are said to be missing not at random (MNAR) if the likelihood that a value of a variable X_1 is missing is related to the missing values of X_1 . When data are MNAR, almost all the standard statistical techniques are no longer applicable (Ref 42). For the available dataset, none of the above-mentioned missing data mechanisms can be ruled out with 100% certainty. Indeed, MCAR is estimated as the most likely missing data mechanism, due to the apparent random disclosure of information about spray parameters by the authors of the academic papers included in the dataset.

Configurations 0-2 do not contain any missing values and thus do not require the application of a null-value replacement strategy. For configuration 3-10, the addition of parameters comes along with the inclusion of null values in the dataset. To handle the missing values in the datasets (configuration 3-10), different null-value replacement strategies were applied.

The complete case analysis handles the missing values by deleting the rows having null values. That means instances with missing values are omitted from the dataset. The advantage of this method is that there is no data replacement required, and therefore, inaccurate data

replacements are avoided. On the other hand, it leads to a loss of information in the dataset. This strategy works best if the data are MCAR, as in this case an analysis using only complete cases will not be biased (Ref 42).

Table 4 shows that from configuration 0-10 more parameters are included into the dataset; indeed, the range of each numerical parameter and the number of types of each categorical parameter are decreasing, leading to a model with a smaller application range.

The mean/mode strategy is a univariate imputation technique, which means that it only uses the column with missing values for imputing values in that column. For numeric data, the mean of each parameter was calculated, and for categorical data the mode for each parameter was determined on the training data set. Missing values in each row of the training data set were then replaced by the mean or mode of the respective parameter. This is an approximation which can increase the model bias. Indeed, the loss of the data can be avoided which can lead to better results compared to the removal of instances with missing values (Ref 42).

K -nearest neighbor (k -NN) imputation algorithms are efficient methods to fill in missing data where each missing value in a data row is replaced by a value obtained from related cases in the whole data set (Ref 43). It belongs to the multivariate imputation techniques. In contrast with the univariate imputation techniques, other variables in the data are factored in to make better predictions about the potential true value of a missing value. K -nearest neighbor imputation algorithms are applicable in any of the three missing data mechanisms, as long as there is a relationship between the variable with the missing value and the other variables (Ref 6).

Table 3 Definition of dataset configuration 0-10 via floating selection of cold spray parameters

	Powder material	Substrate shape	Process gas type	Gas pressure	Spray angle	Gas temperature	Substrate material	Standoff distance	Powder morphology	Traverse speed	Powder feeder rate	Cold spray system	Nozzle type	Avg. particle size
Con. 0	x	x	x	x										
Con. 1	x	x	x	x	x									
Con. 2	x	x	x	x	x	x								
Con. 3	x	x	x	x	x	x	x							
Con. 4	x	x	x	x	x	x	x	x						
Con. 5	x	x	x	x	x	x	x	x	x					
Con. 6	x	x	x	x	x	x	x	x	x	x				
Con. 7	x	x	x	x	x	x	x	x	x	x	x			
Con. 8	x	x	x	x	x	x	x	x	x	x	x	x		
Con. 9	x	x	x	x	x	x	x	x	x	x	x	x	x	
Con. 10	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Avg. = average; Con. = configuration.

Table 4 Parameter range for dataset configuration 0-10 for numerical parameters (top) and categorical parameters (bottom)

Train data sets	Range										
	Con. 0	Con. 1	Con. 2	Con. 3	Con. 4	Con. 5	Con. 6	Con. 7	Con. 8	Con. 9	Con. 10
<i>x</i>											
Numerical parameters:											
Gas pressure [MPa]	0.5-7	0.5-7	0.5-7	0.5-7	0.5-5	0.75-7	0.75-5	0.75-5	0.5-5	0.5-5	0.5-4.14
Spray angle [°]		90-125	90-125	90-125	90-125	90-125	90-125	90-125	90-125	90-125	90
Gas temperature [°]			20-1100	20-1100	20-1100	20-1100	20-1100	20-1000	50-1000	50-1000	50-800
Standoff distance [mm]					10-80	15-80	15-80	20-80	20-80	20-80	25-80
Traverse speed [mm/s]							1.7-3730	1.7-3730	1.7-3730	25-1000	5-1000
Powder feeder rate [g/min]								1.7-120	1.7-120	3.5-48	3.5-20
Avg. particle size [µm]											29-41
Categorical parameters:											
Process gas type	3	3	3	3	3	3	3	3	2	2	2
Powder material	7	7	7	7	7	6	5	5	5	3	2
Substrate shape	2	2	2	2	2	2	2	1	1	1	1
Substrate material				6	6	6	6	6	5	4	3
Powder morphology						8	6	6	2	2	2
Cold spray system									5	3	2
Nozzle type										5	4
<i>y</i>											
Porosity [%]	0-43	0-43	0-43	0-43	0-43	0-43	0-43	0-25.3	0-20.1	0-20.1	0-20.1

Avg. = average; Con. = configuration.

A *k*-NN-imputed value is either a value that was actually measured for another record in a database (1-NN) or the average of measured values from *k* records (*k*-NN) (Ref 43). The algorithm works by calculating the Euclidian distance from the row with the missing value to all the other corresponding row elements (Ref 6). Based on the calculated Euclidian distance, the *k*-NN finds a set of *k* nearest neighbors and then replaces the missing observations for a given parameter by the average of its neighbors' observed values.

Euclidian distance

$$= \sqrt{\text{weight} * (\text{distance from present coordinates})^2}$$

(Eq 2)

where weight: total # of coordinates / # of present coordinates

Multivariate imputation by chained equations (MICE) is a multivariate imputation technique that is used under the assumption that the missing data are MCAR or MAR. Implementing MICE when data are MNAR could result in biased results (Ref 44).

This technique focuses on one missing variable at a time. Once the focus is placed on one variable, MICE uses all the other variables in the data set to predict missing

values in that variable. The prediction is based on a regression model that uses the currently focused missing variable as the dependent variable and all other variables as independent variables (Ref 44).

Handling of Categorical Features

A categorical feature is one that has two or more categories, but there is no intrinsic ordering to the categories. To use categorical features in a dataset to train a model, they first must be transformed into numerical features. Converting categorical data into numerical data can be done by integer encoding. In this procedure, each category of a categorical variable is assigned an integer value, for example, for the categorical variable 'powder material' the category CP titanium is 1 and the category copper is 2, etc. The problem with integer encoding is that the assigned numbers represent a natural ordered relationship, and the machine learning algorithms could assume that higher numbers are more important than lower numbers. Therefore, for categorical variables where no such ordinal relationship exists, the integer encoding is not sufficient (Ref 45).

In one hot encoding (OHE) for each category of a categorical feature, one binary attribute (which can have the

value 0 or 1) is created. Therefore, it can be avoided that the machine learning model assumes a natural ordering between the categories. Table 5 gives an example of the one hot encoded categorical variable 'Powder material'. In this case, instance 1 is a CP titanium sample, instance 2 is a copper sample, and instance 3 is a tantalum sample (Ref 45).

The downside of this method is, that with every category, the feature space gets expanded, increasing the complexity of the model. Here, the OHE feature handling methods was applied. Further, the CatB machine learning algorithm uses the integrated categorical feature handling method OTE (see section machine learning algorithms), which was also used in this study.

Training Data and Testing Data

To separate the existing datasets (configurations 0-10) for the training and testing procedure, the datasets were split into a training and testing portion. Therefore, two different strategies were applied. For the complete case analysis, the datasets were split into training and testing sets in an 80:20 ratio. For the missing data replacement strategies, the data without null values were used as the testing dataset and the data with missing values were used as the training dataset. The latter strategy was selected to allow for a comparison of the applied null value replacement strategies.

Method

The goal of this study was to find the optimal data set configuration, null value replacement strategy, categorical feature handling method, and machine learning algorithm with optimal hyperparameters to achieve the model with the best prediction capacity. Figure 3 shows the examined combinations of these impact factors.

Prediction Accuracy

The root mean squared error (RMSE) is a common performance metric that gives a good idea of the accuracy of a

Table 5 Example for the application of the OHE strategy for the transformation of the categorical features 'powder material' into numerical features

Instance #	Copper	CP titanium	Tantalum
1	0	1	0
2	1	0	0
3	0	0	1

model given the range of the output values. A small RMSE means the model can make predictions with a high accuracy (Ref 46).

It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_p - y)^2} \quad (\text{Eq 3})$$

where y_p : predicted output value for test set input values, y : corresponding output value from test set (target), m : number of test sets.

Several authors reported measurement ranges for the percentage of porosity for each examined cold spray sample (Ref 12, 15, 19, 47). The average of those reported measurement ranges is 1.72% porosity. Therefore, the target accuracy of a machine learning model aiming to predict the porosity in cold spray was chosen to be in a similar range of 2% porosity.

For each setup described in Table 3, a hyperparameter optimization for the respective applied machine learning algorithm was conducted. A "one at a time" approach and a randomized search were applied to find the best hyperparameter combinations for each setup. In both cases, a 10-fold cross-validation based on the RMSE as a performance indicator was performed on the training data set of the respective configuration. The "one at a time" approach is a systematic approach, where every hyperparameter is tackled one at a time, using the best results for subsequent iterations (Ref 48). For the randomized search, which is based on inputting all the hyperparameter ranges at once, the number of iterations was set to 60. With 60 iterations, 95% of the time, the best 5% sets of parameters can be found.

With this approach, more than 6000 models with different setups and hyperparameter combinations were evaluated on the training data. From this pool, the hyperparameter-optimized models for each setup were evaluated using the test data set of the respective data set configuration.

Feature Importance

The data set configuration 0-10 differs due the fact that they include different parameters (see Table 3), which can affect the models performances. To find out about which features have the most prominent effect on the porosity, a feature importance study was conducted. Indeed, it is assumed that the applied data preparation strategies alter the structure of the dataset. Therefore, an analysis of the effect of the impact factors mentioned in Fig. 3 on the feature importance was also conducted. Here, the Shapley value approach was selected for this analysis.

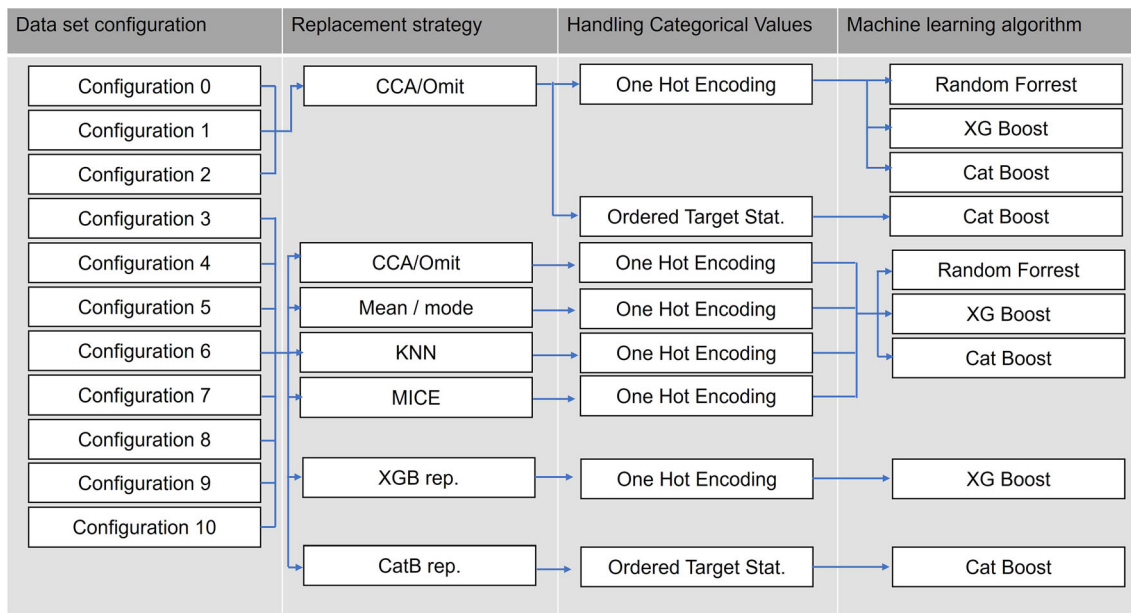


Fig. 3 Overview of the dataset configurations, applied data preparation strategies and machine learning algorithms tested

The concept of Shapley values was originally developed in Game Theory and is applied to estimate the contribution of each player to the output of a game. The Shapley value for a player is determined by calculating the expected marginal contribution of the player to the game’s output (i.e., value). The expected marginal contribution is defined as the average contribution of a player across all possible permutations or coalitions of players. All the marginal contributions of a player are weighted by the probabilities that they make the contribution. The weighted contributions are summed over all the coalitions that the player can join. Therefore, a player’s individual contribution and the interactions between players are considered (Ref 49, 50).

In ML Shapley values can be used to gain an understanding of the relationship between the prediction of the model and the components of the data instance that the model used to generate that prediction. Therefore, the value of the game is the model prediction and the feature values are the players (Ref 51).

The Shapley value Φ_i for the local importance (i.e., influence of a feature on the prediction made by a model for a specific instance) of the i th feature can be calculated as follows (Ref 51):

$$\Phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{Eq 4}$$

where $|S|$: size of the subset before i th feature is added, $|F|$: number of features, $f_{S \cup \{i\}}(x_{S \cup \{i\}})$: prediction made by the model with i th feature is included, $f_S(x_S)$: prediction made

by the model with i th feature is NOT included, $\frac{|S|!(|F| - |S| - 1)!}{|F|!}$: the weight for combinations for this occurrence, $S \subseteq F \setminus \{i\}$: all possible subsets without i th feature, $S \cup \{i\}$: a subset with i th feature added.

To calculate the global feature importance over the entire model, the average of the absolute Shapley values for all the instances (i.e., the local feature importance for each instance) is calculated (Ref 51).

When interpreting the Shapley values, it must be considered that they are not a measure of how important a given feature is in the real world. It is a measure of how important a feature is to the model. A model is not necessarily a good representation of reality, as predictions can be incorrect. The Shapley values essentially give the contributions of the model features to a prediction that usually deviates from the target variable. Therefore, no conclusion that goes beyond the model should be made (Ref 51).

Null Value Replacement Accuracy

Univariate and multivariate null value replacement techniques have been applied to handle missing values in the dataset. Subsets consisting of only complete instances of the data configurations 3-10 were used to assess the different null value replacement strategies. Therefore, null values were randomly inserted into these datasets. The percentage of null values was chosen to be equivalent to the percentage of null values in the dataset configurations that were used to train the machine learning models. The univariate replacement strategies did not necessitate the conversion of the categorical variables into numerical

variables. Indeed, for the application of the multivariate null value replacement strategies a conversion of the categorical features into numerical features was required. For the conversion, the OHE strategy was used. Consequently, the percentage of data completeness varies depending on which type of null value replacement strategy was applied. Table 6 shows the percentage of completeness for both the univariate replacement strategy and for multivariate replacement strategies datasets.

The random insertion of null values was repeated ten times for each dataset to allow for a statistically relevant assessment of the replacement strategies. Subsequently, the univariate (mean, mode) and multivariate (KNN and MICE) replacement strategies were applied on each dataset to replace the missing values. Due to the different scales of the parameters of the datasets, a standard scaler was applied to each value of the datasets (original complete datasets and datasets with replaced missing values) scaling each parameter to unit variance. The standard score of a sample x is calculated as (Ref 52):

$$z = (x - \mu) / s \quad (\text{Eq 5})$$

where μ : feature mean, s : feature standard deviation.

With standardized values in the dataset, a comparison of the different applied null value replacement strategies for each dataset configuration 3-10 using the RMSE was possible. In this case, the RMSE was calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (\text{inserted value} - \text{true value})^2} \quad (\text{Eq 6})$$

where m : number of missing values.

Results and Discussion

In this section, the result of the application of the methods introduced in “Method” section is discussed. An evaluation of the hyperparameter optimized models of each setup was performed on the test dataset. The features which have the biggest effect on the porosity in cold spray were identified with the help of the Shapley value approach. The accuracies of the applied null value replacement strategies were determined. Finally, limitations of the approach and suggestions of future works are discussed.

Evaluation of Hyperparameter-Optimized Machine Learning Models

For each setup in Fig. 3 a hyperparameter optimization was performed as described in chapter 4.1, followed by an evaluation of the hyperparameter optimized models using the test dataset. Figure 4 gives an overview of the performance of all the hyperparameter-optimized models on the test dataset using the RMSE as a performance indicator.

Figure 4 shows that most of the model’s predictions deviate 3-4% porosity on average (denoted by the RMSE class (3,4]) from the true percentage porosity values. Indeed, 12 of the developed models are in the target range of 0-2% porosity RMSE of their predictions and are therefore within the specified target prediction accuracy. This means that these models can predict the percentage of porosity in a cold-sprayed part with a deviation of 0-2% porosity on average.

Table 7 shows the setup of these models and their performance on the testing data and the training data in detail.

In most of the cases, the test error is less than the training error indicating a possible sampling bias in the test. This means that the respective test dataset possibly does not reflect the realities of the environment in which the model was trained. On the other hand, a much better performance on the training set as compared to the testing set would indicate an overfitting of the model to the training data. Comparing the results for the complete case analyses and the application of null value replacement strategies, it must be considered that for the former the training and testing set size decreased with the inclusion of parameters to the dataset (Table 8). This dataset shrinkage only occurred for the complete case analysis as in this case instances with missing values were removed from the dataset.

This dataset shrinkage must be considered when evaluating the prediction accuracy of the models on the testing dataset. The evaluation of the models using the testing datasets with a higher number of testing samples provides more reliable results, whereas the evaluation of the models with a very small test dataset is less conclusive. Therefore, the use of a model that was tested with a larger test dataset is preferred in the case of similar evaluation results. It is noticed that all models for which the complete case strategy was applied were trained with a small dataset (19-52 samples) and therefore tested with a small dataset (5-13

Table 6 Null value percentage of dataset configurations 3-10 for univariate and multivariate replacement strategies

Strategy	Con. 3	Con. 4	Con. 5	Con. 6	Con. 7	Con. 8	Con. 9	Con. 10
Univariate	0.99	0.99	0.98	0.95	0.92	0.89	0.85	0.82
Multivariate	0.99	0.99	0.95	0.95	0.94	0.88	0.78	0.77

Con. = configuration.

samples) as well. On the other hand, it is noticeable that models trained with a dataset prepared using a missing value replacement strategy have a lower number of parameters for each dataset. As can be seen from Table 6, these datasets have lower percentages of missing values. Therefore, there is reason to believe that the missing handling replacement strategies were effective only when dealing with a low percentage of missing values.

Figure 5 presents a comparative analysis of predicted and actual porosity values derived from the respective test datasets for two models, one employing a complete case strategy (Table 7 Nr. 3) and the other utilizing a null value replacement strategy (Table 7 Nr. 7). The actual porosity values are the real measurements of porosity taken from the test samples, serving as the true reference for porosity in this analysis. The closer the points are to the red 45° line in each plot, the more accurate was the prediction of the respective model. The predictions of both models come close to the actual porosity values of the test samples.

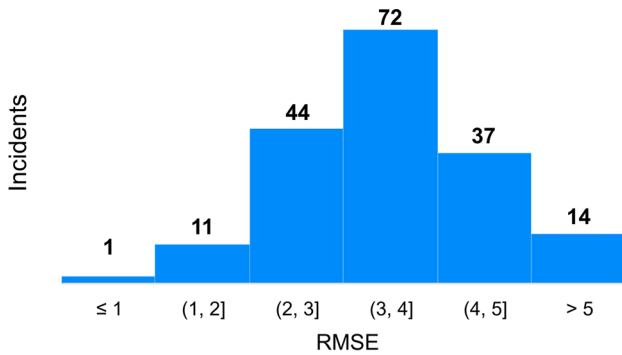


Fig. 4 Performance of hyperparameter-optimized models for each setup using the RMSE as a performance indicator

Indeed, in Fig. 5(a) a large evaluation gap in the test dataset can be observed between ca. 4% and 12.5% porosity. This may be due to the small test set size used for models trained with a dataset with many parameters that was prepared using the complete case strategy. Figure 5(b) does not exhibit an as big evaluation gap, due to the larger test set that was used to evaluate the model trained with the dataset for which a missing value replacement strategy was applied.

Table 9 shows the average deviation of predicted and real porosity values of the test data samples in different porosity ranges. Model #7 performs best in the porosity region from 1 to 10%, whereas the predictions for parts with a high porosity deviate more significantly from the real porosity values. Model #3 performs best in the porosity

Table 8 Change of training and testing set size depending on the dataset configuration

Dataset Con.	Train data	Test data
0	193	49
1	193	49
2	193	49
3	188	48
4	180	45
5	155	39
6	116	30
7	91	23
8	52	13
9	42	11
10	19	5

Con. = configuration.

Table 7 Overview of all hyperparameter-optimized models with a RMSE of < 2%

#	Data set con.	Null value replacement	Categorical feature handling	ML algorithm	RMSE testing data	RMSE training data
1	8	Complete case	OHE	XGB	1.61	0.83
2	10	Complete case	OHE	XGB	1.54	3.11
3	8	Complete case	OHE	CatB	1.34	0.49
4	9	Complete case	OHE	CatB	1.77	0.61
5	10	Complete case	OHE	CatB	1.65	0.01
6	1	Univariate	OHE	CatB	1.79	2.24
7	2	Univariate	OHE	CatB	1.61	1.73
8	0	Multivariate—KNN	OHE	CatB	1.82	2.61
9	4	Multivariate—MICE	OHE	CatB	1.98	2.64
10	9	Complete case	OHE	CatB	1.78	2.07
11	10	Complete case	OHE	CatB	1.57	3.10
12	10	Complete case	OHE	RF	0.79	2.77

Con. = configuration.

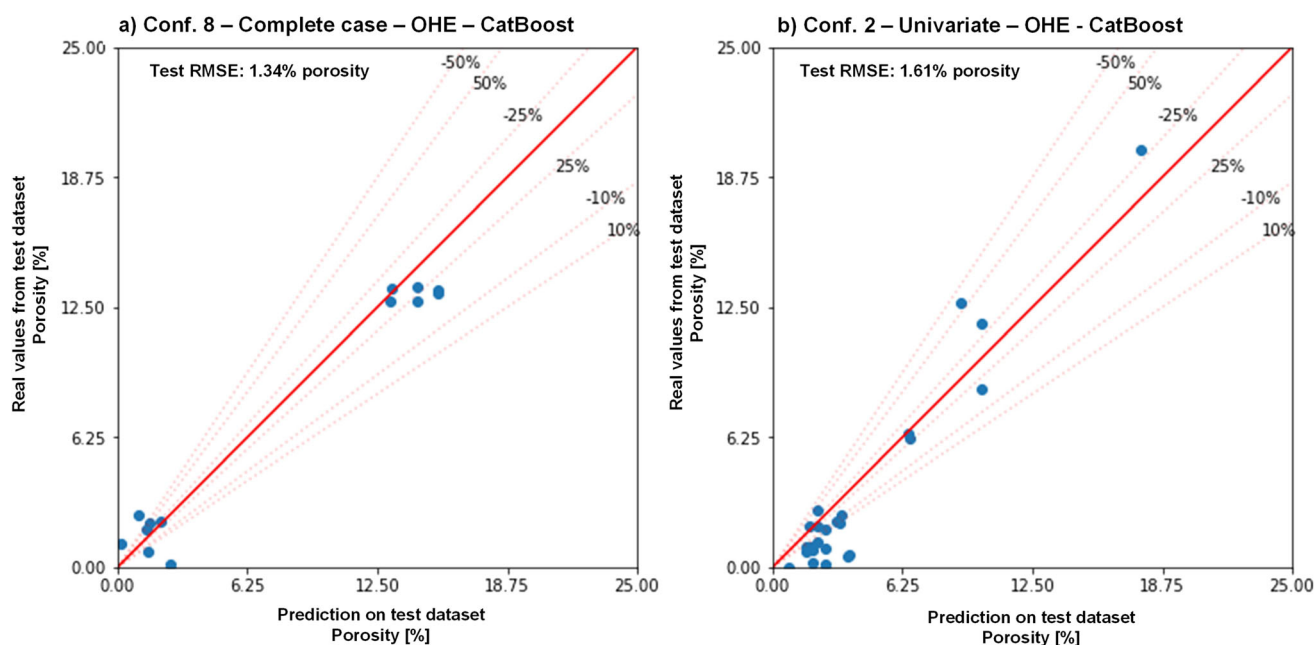


Fig. 5 Comparison of predicted target values versus actual target values. **a** Conf. 8—complete case—OHE—CatB, **b** Conf. 2—Univariate—OHE—CatB

Table 9 Differences between predicted target values and actual target values for testing portion per bin

#	Data set con.	Null value replacement	Categorical feature handling	ML algorithm	Bin-wise performance				
					[0-1%]	(1-2%]	(2-3%]	(3-10%]	> 10%
3	8	Complete Case	OHE	CatB	1.60	0.70	0.74	...	1.23
7	2	Univariate	OHE	CatB	1.60	0.50	0.83	0.66	2.55

region from 1 to 3%. Due to the small test set size, there were no samples in the porosity region from 3 to 10%; therefore, the performance of model #3 in this region could not be evaluated.

Feature Importance Analysis Using the Shapley Value Approach

The Shapley value approach was applied to test the effect of different dataset configurations on the feature importance. Figure 6 shows the Shapley values of every feature for every sample for a CatB algorithm trained with a configuration 10 dataset. It uses the Shapley values to show the distribution of the impacts each feature has on the model output. The color represents the feature value (red: high, blue: low, gray: categorical). For example, low values of the feature ‘Gas pressure’ demonstrate a strong positive (increasing) effect on the resulting porosity, whereas high values of this feature have a negative (decreasing) effect on the porosity.

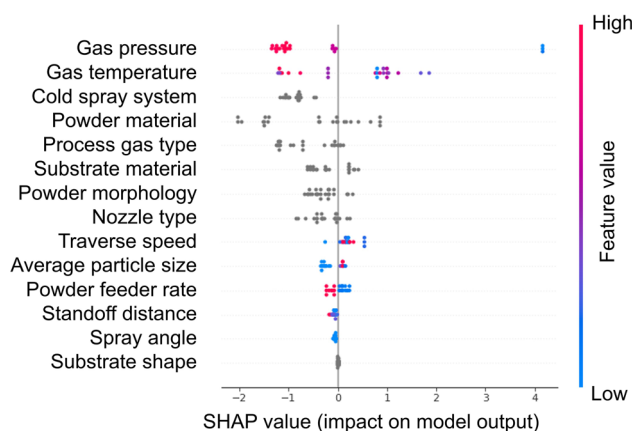


Fig. 6 Shapley value distribution of a CatB algorithm trained with a configuration 10 dataset

Calculating the Shapley value for a specific feature for every sample in the dataset and averaging the absolute values lead to the feature importance of the entire model. Figure 7 shows the mean absolute Shapley value calculated

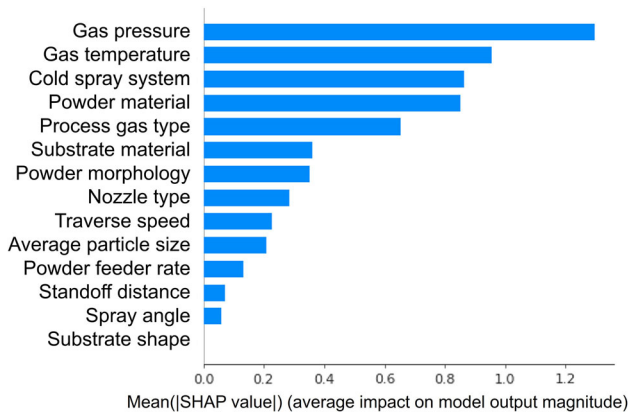


Fig. 7 Mean absolute Shapley values of all features of a CatB algorithm trained with a configuration 10 dataset

for every sample and every feature of a CatB algorithm trained with the configuration 10 dataset.

Therefore, for this model it can be concluded that the most important features in the model are the gas pressure, gas temperature and the Cold spray system, whereas less important features are the nozzle standoff distance, the spray angle, and the substrate shape. This approach was taken a step further by calculating the average of the mean absolute Shapley values of all tested models. Table 10 shows the average of the mean absolute Shapley values for every feature.

A comparison of the average mean absolute Shapley values from Table 10 with the mean absolute Shapley values of the CatB model trained with the configuration 10 dataset (Fig. 7) shows that there are differences in the importance of single features across the different models. For instance, the feature cold spray system is the third most important feature for the CatB model trained on the configuration 10 dataset, whereas this feature is the third least important feature across all the tested models. Table 11 summarizes the average mean absolute Shapley values depending on the applied machine learning algorithm. The lower part of the table shows a ranking of the features depending on the respective average mean absolute Shapley value. As a conclusion, there are only minor differences in how the choice of the machine learning algorithm affects the importance of the single features.

Table 12 summarizes the difference of the average mean absolute Shapley values as per different applied null value replacement strategy. When considering the feature rankings in the bottom half of the table, it appears that there are only minor differences between the univariate (Mean, mode) and the multivariate (KNN, MICE) replacement strategies. However, there is a major difference evident between those strategies, the complete case analysis, and the XGB strategy, for the features of spray angle, powder feeder rate, and particle size. Therefore, the selection of the

Table 10 Average mean absolute Shapley values across all models

Feature	Shapley value	
	Average	STD
Gas temperature	2.39	0.86
Gas pressure	1.40	0.55
Avg. particle size	1.18	1.05
Traverse speed	0.61	0.57
Process gas type	0.53	0.51
Powder feeder rate	0.51	0.44
Standoff distance	0.32	0.24
Substrate material	0.22	0.27
Spray angle	0.21	0.29
Powder material	0.19	0.38
Powder morphology	0.13	0.25
Cold spray system	0.11	0.28
Nozzle type	0.04	0.10
Substrate shape	0.00	0.00

Avg. = average.

null value replacement strategy significantly affects the resulting feature importance of the models.

Table 13 compares the two different applied categorical feature handling strategies—OHE and OTE—for a complete case analysis and a univariate replacement strategy. In both cases, the complete case analysis and the univariate replacement strategy, it is apparent that the categorical feature strategy also has a major impact on the determined feature importance of the respective models. Major differences can be observed for the features powder material, spray angle, standoff distance, traverse speed, powder feeder rate, cold spray system, nozzle type, and average particle size.

Altogether, it is apparent that drawing a conclusion from the Shapley values to the effect of a feature on the porosity is not advisable, as the applied data processing steps (null value replacement strategy and categorical feature handling) have a major impact on the structure of the data and the Shapley values.

Indeed, for the features of substrate shape and gas temperature a similar ranking was observed across all different analyses. The substrate shape is constantly ranked as a feature with a low impact on the porosity, whereas the gas temperature is constantly ranked as a highly important feature. The feature substrate shape is a two-category feature for which the feature category “flat” (240 instances) was highly overrepresented compared to the feature category “round” (two instances) which is the reason for the low impact of this feature on the porosity. Therefore, in any case the developed models should only be applied for

Table 11 Comparison of the average mean absolute Shapley values for each applied machine learning algorithm

Shapley value	ML algorithm						Avg.
	XGB		RF		CatB		
	Avg.	Std.	Avg.	Std.	Avg.	Std.	
Powder material	0.19	0.36	0.17	0.32	0.21	0.75	0.19
Substrate shape	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Process gas type	0.57	0.52	0.55	0.54	0.48	0.48	0.53
Substrate material	0.24	0.30	0.18	0.19	0.24	0.28	0.22
Gas pressure	1.52	0.56	1.25	0.47	1.42	0.58	1.39
Spray angle	0.22	0.20	0.23	0.45	0.18	0.18	0.21
Gas temperature	2.65	0.66	2.87	0.65	1.80	0.82	2.44
Standoff distance	0.43	0.28	0.24	0.15	0.29	0.24	0.32
Powder morphology	0.15	0.29	0.09	0.27	0.14	0.26	0.13
Traverse speed	0.86	0.71	0.57	0.56	0.42	0.27	0.62
Powder feeder rate	0.55	0.35	0.65	0.64	0.37	0.23	0.53
Cold spray system	0.12	0.19	0.07	0.26	0.15	0.27	0.11
Nozzle type	0.03	0.05	0.04	0.09	0.06	0.14	0.04
Avg. particle size	1.47	1.20	1.20	1.01	0.92	0.84	1.20
Rank	XGB		RF		CatB		Avg. rank
Powder material	10		10		9		10
Substrate shape	14		14		14		14
Process gas type	5		6		4		5
Substrate material	8		9		8		8
Gas pressure	2		2		2		2
Spray angle	9		8		10		9
Gas temperature	1		1		1		1
Standoff distance	7		7		7		7
Powder morphology	11		11		12		11
Traverse speed	4		5		5		5
Powder feeder rate	6		4		6		5
Cold spray system	12		12		11		12
Nozzle type	13		13		13		13
Avg. particle size	3		3		3		3

Avg. = average; Std. = standard deviation.

the porosity prediction of “flat” sprays. The high significance of the gas temperature on the porosity is in line with experimental studies (Ref 17) and can, due to consistent results across all models, be assumed as valid.

Evaluation of Different Null Value Replacement Strategies

Following the description of the method in “Null Value Replacement Accuracy” section, an evaluation of the null value replacement strategies was conducted. Table 14 gives

an overview of the accuracy of the null value replacements of each applied strategy for each dataset configuration expressed through the RMSE.

From those results, it can be concluded that the Univariate replacement strategy and the KNN-replacement strategy led to comparable results in terms of accurately replacing missing values. The clearly best result was achieved using the MICE replacement strategy on dataset configurations with a lower number of parameters (dataset configurations 3 and 4).

Table 15 gives an overview of the RMSE for the MICE replacement strategy applied on dataset configuration 3 without the application of the standard scaler for different parameters. Without the standard scaler, an interpretation of the deviation between the true values and the imputed values for each parameter is possible.

On the categorical parameters and the numerical parameter ‘spray angle,’ the MICE strategy performed well. Indeed, the performance on the numerical parameters ‘Gas temperature’ and ‘Gas pressure’ is unsatisfactory.

Limitations of the Approach and Future Works

A major downside of the reported approach is the limited control over the available data. For instance, the input parameters particle velocity and particle temperature are two main factors that affect the porosity level of a cold spray deposit [21, 53]. These parameters were not included in the development of the models reported here. This is because these parameters were not measured and reported together with the porosity in a sufficient quantity of reports. A possible solution could be the development of a model for the prediction of the particle velocity and/or particle temperature that could be used as an intermediate model to the existing porosity prediction models. This indeed is out of the scope of this work.

This study explores various dataset configurations with different input parameters. Configurations with fewer parameters, such as configuration 4, have the drawback of not considering variations in parameters excluded from the dataset, such as powder morphology, traverse speed, powder feeder rate, cold spray system, nozzle type, and average particle size (refer to Table 3). This may lead the model to learn incorrect relationships between input and output data.

On the contrary, certain input parameters exert a more pronounced impact on porosity than others. Removing irrelevant or redundant features from the dataset, thereby reducing model complexity, serves as a strategy to counteract overfitting. Hence, in this study, different dataset configurations were tested and evaluated to find a good balance between model complexity and overfitting.

Table 12 Comparison of the average mean absolute Shapley values for each applied null value replacement strategy

Shapley value	Null replacement strategy										Avg.
	Complete Case analysis		Mean-mode		KNN		MICE		XGB strategy		
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.	
Powder material	0.27	0.48	0.15	0.32	0.14	0.25	0.13	0.29	0.25	0.45	0.16
Substrate shape	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Process gas type	0.31	0.30	0.61	0.56	0.62	0.56	0.51	0.55	0.41	0.35	0.41
Substrate material	0.23	0.38	0.20	0.21	0.22	0.21	0.20	0.21	0.26	0.34	0.18
Gas pressure	1.22	0.87	1.37	0.32	1.42	0.26	1.37	0.46	1.88	0.56	1.21
Spray angle	0.55	0.50	0.12	0.14	0.10	0.11	0.10	0.11	0.33	0.14	0.20
Gas temperature	1.75	0.94	2.99	0.52	2.73	0.50	2.66	0.46	1.96	0.20	2.01
Standoff distance	0.38	0.34	0.45	0.30	0.29	0.18	0.28	0.26	0.52	0.33	0.32
Powder morphology	0.10	0.18	0.07	0.14	0.12	0.24	0.15	0.26	0.17	0.25	0.10
Traverse speed	0.82	1.21	0.71	0.26	0.34	0.13	0.60	0.29	1.23	0.54	0.62
Powder feeder rate	0.44	0.37	0.44	0.25	0.43	0.15	0.66	0.28	0.29	0.09	0.38
Cold spray system	0.23	0.65	0.05	0.12	0.05	0.07	0.17	0.30	0.20	0.28	0.12
Nozzle type	0.06	0.22	0.02	0.03	0.04	0.07	0.07	0.10	0.03	0.06	0.04
Avg. particle size	0.03	0.03	0.93	0.22	2.68	0.46	2.17	0.33	0.40	0.00	1.03

Rank	Complete case analysis	Mean-mode	KNN	MICE	XGB strategy	Avg. rank
Powder material	8	9	9	11	10	9
Substrate shape	14	14	14	14	14	14
Process gas type	7	5	4	6	5	5
Substrate material	10	8	8	8	9	9
Gas pressure	2	2	3	3	2	2
Spray angle	4	10	11	12	7	9
Gas temperature	1	1	1	1	1	1
Standoff distance	6	6	7	7	4	6
Powder morphology	11	11	10	10	12	11
Traverse speed	3	4	6	5	3	4
Powder feeder rate	5	7	5	4	8	6
Cold spray system	9	12	12	9	11	11
Nozzle type	12	13	13	13	13	13
Avg. particle size	13	3	2	2	6	5

Avg. = average; Std. = standard deviation.

The most effective solution would be to generate a firsthand dataset with comprehensive knowledge of the experimental procedure. Generating a first-hand dataset and employing it for the development of a predictive ML model could be viable if a smaller set of parameters was varied. While this would decrease the effort required for data generation, it would also limit the model’s applicability to diverse spray scenarios.

Another point for discussion involves the porosity levels reported in the papers, which were measured using two different techniques. 90% of the porosity samples were measured using image analysis following a metallographic polishing of cross sections of the samples. With this

method, factors such as inconsistent sample preparation, inconsistent image quality, choice of thresholding and filtering method, and subjectivity in selecting regions of interest can contribute to inconsistent results. The other 10% of the samples were measured using the Archimedes method. Here the results can be affected by inconsistent surface cleaning of the samples, incomplete penetration of the liquid into the pores, and insufficient drying of the samples. Therefore, both methods are highly susceptible to operator bias. Altogether, this is expected to reduce the consistency of the porosity measurement results in the dataset and therefore negatively affects the prediction accuracy of the models trained with the dataset. This is a

Table 13 Comparison of the average mean absolute Shapley values for each applied feature handling strategies

Shapley value	Complete case analysis				Univariate replacement strategy			
	OHE		OTE		OHE		OTE	
	Avg.	Std.	Avg.	Std.	Avg.	Std.	Avg.	Std.
Powder material	0.23	0.39	1.51	0.90	0.15	0.40	1.12	0.36
Substrate shape	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Process gas type	0.27	0.25	0.50	0.28	0.53	0.46	0.54	0.18
Substrate material	0.23	0.32	1.45	0.68	0.22	0.22	0.68	0.31
Gas pressure	1.17	0.78	0.31	0.19	1.32	0.40	1.54	0.55
Spray angle	0.35	0.24	0.73	0.36	0.17	0.15	0.12	0.12
Gas temperature	1.20	0.58	1.16	0.43	2.53	0.48	0.96	0.41
Standoff distance	0.53	0.38	0.25	0.17	0.32	0.20	0.10	0.06
Powder morphology	0.12	0.15	0.79	0.59	0.07	0.11	0.43	0.17
Traverse speed	0.64	0.34	0.29	0.23	0.43	0.09	0.29	0.18
Powder feeder rate	0.63	0.29	0.28	0.15	0.42	0.16	0.17	0.11
Cold spray system	0.11	0.22	0.83	0.68	0.04	0.11	0.85	0.10
Nozzle type	0.14	0.36	0.21	0.13	0.02	0.03	0.34	0.06
Avg. particle size	0.01	0.00	0.15	0.00	0.70	0.00	0.21	0.00
Rank	Complete case analysis OHE		Complete case analysis OTE		Univariate replacement strategy OHE		Univariate replacement strategy OTE	
Powder material	8		1		10		2	
Substrate shape	14		14		14		14	
Process gas type	7		7		4		6	
Substrate material	9		2		8		5	
Gas pressure	2		8		2		1	
Spray angle	6		6		9		12	
Gas temperature	1		3		1		3	
Standoff distance	5		11		7		13	
Powder morphology	11		5		11		7	
Traverse speed	3		9		5		9	
Powder feeder rate	4		10		6		11	
Cold spray system	12		4		12		4	
Nozzle type	10		12		13		8	
Avg. particle size	13		13		3		10	

Avg. = average.

disadvantage that can only be truly overcome by developing a model that was trained by a dataset that was generated by one person or a small group of people using the exact same method. One drawback of this practice indeed is that generating an equivalent volume of data would be significantly more challenging.

Due to the presence of categorical features, such as process gas type, powder, and substrate material type, etc., different categorical feature replacement strategies have been applied in this study. Another way of transforming the categorical values into numerical values would have been to use material properties such as density, yield strength, or melting point. These parameters are proved to indirectly

Table 14 Comparison of the accuracy of the null value replacement strategies applied on standard scaled datasets using the RMSE

Dataset configuration	RMSE Univariate replacement	RMSE KNN replacement	RMSE MICE replacement
3	1.03	0.84	0.49
4	1.05	1.06	0.51
5	0.96	1.01	0.76
6	1.03	1.00	0.85
7	0.73	1.03	0.94
8	0.83	0.85	0.78
9	0.98	0.89	0.79
10	1.05	0.94	0.78

Table 15 Application of the MICE replacement strategy for a configuration 3 dataset—without the application of a standard scaler

	Categorical parameters	Gas pressure, MPa	Gas temperature, °C	Spray angle, °
RMSE	0.13	0.93	130.58	0.0
Number of replaced values	89	5	4	4a

affect the porosity through their influence on the critical velocity and actual particle velocity (Ref 54-56). Indeed, the values for those parameters were rarely measured or reported together with the porosity. Therefore, this approach was rejected.

Finally, it is crucial to ensure that any input data used with the trained models fall within the ranges of the training data to achieve a prediction accuracy similar to that demonstrated in the manuscript. This requirement applies to both directly included parameters in the model and those indirectly influencing it. For example, a model trained with dataset configuration 10 includes particle size as an input parameter and will adjust its output based on changes in the particle size input parameter, among others. To achieve an accuracy comparable to that reported in the manuscript, the model should only be applied to input parameters within the range of the training dataset.

For a model trained with a configuration 9 dataset, the particle size is not directly included as a parameter. Therefore, it will not respond to a change of the particle size in the input parameters. Indeed, the particle size should also only be varied within the value range of the training set. Selecting an average particle size outside this range may result in a prediction accuracy different from that stated in the manuscript.

Conclusion

In this study, a dataset was generated from the information provided in cold spray academic papers regarding process parameters and the associated part properties. Due to the

structure of the available data, several data pre-processing methods were applied to prepare the data for the training and testing of machine learning algorithms. The goal was to develop a machine learning model that can accurately predict the porosity of a cold-sprayed part with an average prediction error of less than 2% porosity when applied within the input parameter ranges of the training data. The assessment of multiple model setups led to the following conclusions:

- (1) The application of null value replacement strategies and categorical feature handling strategies presumably affects the structure of the dataset and the importance of single features for the predictions of the models.
- (2) In all tested models, the feature gas temperature was ranked as a highly influential feature for the porosity prediction of a cold-sprayed part.
- (3) The tested null value replacement strategies fail to accurately replace all the missing values in the datasets. In the best case with the MISE strategy, the average deviation of replaced missing values for ‘Gas pressure’ and ‘Gas temperature’ to the true values was 0.93 MPa and 130.58 °C, respectively. This is still a high value given that the range for both parameters for the given dataset is 6.5 MPa and 1080 °C, respectively.
- (4) With 72 out of 180 models evaluated with the testing dataset, an average deviation from the predicted porosity to the true porosity of 3-4% was achieved. This is a sufficient accuracy if the model is used to “classify” parts into dense and porous parts. Indeed,

the accuracy would be insufficient to distinguish between fully dense materials.

- (5) With 12 of the models evaluated using a testing dataset, the targeted prediction accuracy of less than 2% porosity prediction error was successfully achieved. Indeed, 8 out of those 12 models were developed using only the complete instances in the dataset. It must be considered that the validation of these models against the testing dataset becomes less conclusive due to the occurring shrinkage of the testing datasets.
- (6) 4 out of the 12 models that achieved an RMSE of less than 2% were trained with a dataset prepared using missing value replacement strategies. The fact that those models were also trained with datasets with a low number of parameters and therefore a low number of missing values led to the conclusion that the application of the applied missing value replacement strategies is only effective in the case of a low percentage of missing values in the dataset.

Acknowledgments This work is supported under a Swinburne University Postgraduate Research Award. The authors acknowledge the support from the Australian Research Council (ARC). The ARC Training Centre in Surface Engineering for Advanced Materials, SEAM, has been funded under the ARC Industrial Transformation Training Centre (ITTC) scheme via Award IC180100005. We are grateful for the additional support of the industrial, university, and other organization partners who have contributed to the establishment and support of SEAM.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. S. Yin, P. Cavaliere, B. Aldwell, R. Jenkins, H. Liao, W. Li, and R. Lupoi, Cold spray additive manufacturing and repair: fundamentals and applications, *Addit. Manuf.*, 2018, **21**, p 628-650.
2. T. Pereira, J.V. Kennedy, and J. Potgieter, A comparison of traditional manufacturing vs additive manufacturing, the best method for the job, *Procedia Manuf.*, 2019, **30**, p 11-18.
3. J. Kruth, P. Mercelis, J. Van Vaerenbergh, L. Froyen, and M. Rombouts, Binding mechanisms in selective laser sintering and selective laser melting, *Rapid Prototyp. J.*, 2005, **11**(1), p 26-36.
4. J. Sun, Y. Han, and K. Cui, Innovative fabrication of porous titanium coating on titanium by cold spraying and vacuum sintering, *Mater. Lett.*, 2008, **62**(21-22), p 3623-3625.
5. H. Wu, X. Xie, M. Liu, C. Verdy, Y. Zhang, H. Liao, and S. Deng, Stable layer-building strategy to enhance cold-spray-based additive manufacturing, *Addit. Manuf.*, 2020, **35**, p 101356.
6. P. Jönsson and C. Wohlin, An evaluation of K-nearest neighbour imputation using likert data, in *Proceedings of the 10th International Symposium on Software Metrics (IEEE, Chicago, IL, USA, 2004)*, pp. 108-118. <https://doi.org/10.1109/METRIC.2004.1357895>
7. A. Choudhuri, P.S. Mohanty, and J. Karthikeyan, Bio-ceramic composite coatings by cold spray technology, in *International Thermal Spray Conference 2009*, (Las Vegas, NV, 2009), pp. 391-396.
8. W. Wong, E. Irissou, A.N. Ryabinin, J.-G. Legoux, and S. Yue, Influence of helium and nitrogen gases on the properties of cold gas dynamic sprayed pure titanium coatings, *J. Therm. Spray Technol.*, 2011, **20**(1-2), p 213-226.
9. A. Vargas-Uscategui, P.C. King, S. Yang, C. Chu, and J. Li, Toolpath planning for cold spray additively manufactured titanium walls and corners: effect on geometry and porosity, *J. Mater. Process. Technol.*, 2021, **298**, p 117272.
10. A.W.-Y. Tan, W. Sun, Y.P. Phang, M. Dai, I. Marinescu, Z. Dong, and E. Liu, Effects of traverse scanning speed of spray nozzle on the microstructure and mechanical properties of cold-sprayed Ti6Al4V coatings, *J. Therm. Spray Technol.*, 2017, **26**(7), p 1484-1497.
11. W. Wong, P. Vo, E. Irissou, A.N. Ryabinin, J.-G. Legoux, and S. Yue, Effect of particle morphology and size distribution on cold-sprayed pure titanium coatings, *J. Therm. Spray Technol.*, 2013, **22**(7), p 1140-1153.
12. S. Yin, P. He, H. Liao, and X. Wang, Deposition features of Ti coating using irregular powders in cold spray, *J. Therm. Spray Technol.*, 2014, **23**(6), p 984-990.
13. X. Song, K.L. Ng, J.M.-K. Chea, W. Sun, A.W.-Y. Tan, W. Zhai, F. Li, I. Marinescu, and E. Liu, Coupled Eulerian-Lagrangian (CEL) simulation of multiple particle impact during metal cold spray process for coating porosity prediction, *Surf. Coat. Technol.*, 2020, **385**, p 125433.
14. O.C. Ozdemir, C.A. Widener, M.J. Carter, and K.W. Johnson, Predicting the effects of powder feeding rates on particle impact conditions and cold spray deposited coatings, *J. Therm. Spray Technol.*, 2017, **26**(7), p 1598-1615.
15. S.H. Zahiri, C.I. Antonio, and M. Jahedi, Elimination of porosity in directly fabricated titanium via cold gas dynamic spraying, *J. Mater. Process. Technol.*, 2009, **209**(2), p 922-929.
16. A. Hamweendo, P.A.I. Popoola, and I. Botef, Mathematical model for predicting process parameters in cold spray of porous Ti coatings, in *Proceedings of the 1st International Conference on Mathematical Methods and Computational Techniques in Science and Engineering*, (Athens, 2014), pp. 225-229.

17. X. Meng, J. Zhang, J. Zhao, Y. Liang, and Y. Zhang, Influence of gas temperature on microstructure and properties of cold spray 304SS coating, *J. Mater. Sci. Technol.*, 2011, **27**(9), p 809-815.
18. S.H. Zahiri, D. Fraser, S. Gulizia, and M. Jahedi, Effect of processing conditions on porosity formation in cold gas dynamic spraying of copper, *J. Therm. Spray Technol.*, 2006, **15**(3), p 422-430.
19. T. Marrocco, D.G. McCartney, P.H. Shipway, and A.J. Sturgeon, Production of titanium deposits by cold-gas dynamic spray: numerical modeling and experimental characterization, *J. Therm. Spray Technol.*, 2006, **15**(2), p 263-272.
20. K. Spencer and M.-X. Zhang, Optimisation of stainless steel cold spray coatings using mixed particle size distributions, *Surf. Coat. Technol.*, 2011, **205**(21-22), p 5135-5140.
21. K. Binder, J. Gottschalk, M. Kollenda, F. Gärtner, and T. Klassen, Influence of impact angle and gas temperature on mechanical properties of titanium cold spray deposits, *J. Therm. Spray Technol.*, 2011, **20**(1-2), p 234-242.
22. P. Richer, B. Jodoin, L. Ajdelsztajn, and E.J. Lavernia, Substrate roughness and thickness effects on cold spray nanocrystalline Al-Mg coatings, *J. Therm. Spray Technol.*, 2006, **15**(2), p 246-254.
23. W. Li, H. Liao, and H. Wang, in *Cold Spraying of Light Alloys*, ed. by H. Dong. Surface Engineering of Light Alloys: Aluminium, Magnesium and Titanium Alloy (Elsevier, New York, 2010), pp. 242-293. <https://doi.org/10.1533/9781845699451.2.242>
24. P. Magarò, A.L. Marino, A. Di Schino, F. Furguele, C. Maletta, R. Pileggi, E. Sgambitterra, C. Testani, and M. Tului, Effect of process parameters on the properties of stellite-6 coatings deposited by cold gas dynamic spray, *Surf. Coat. Technol.*, 2019, **377**, p 124934.
25. S. Weiller and F. Delloro, A numerical study of pore formation mechanisms in aluminium cold spray coatings, *Addit. Manuf.*, 2022, **60**, p 103193.
26. M. Terrone, A. Ardeshiri Lordejani, J. Kondas, and S. Bagherifard, A numerical approach to design and develop freestanding porous structures through cold spray multi-material deposition, *Surf. Coat. Technol.*, 2021, **421**, p 127423.
27. Y.-F. Shi, Z.-X. Yang, S. Ma, P.-L. Kang, C. Shang, P. Hu, and Z.-P. Liu, Machine learning for chemistry: basics and applications, *Engineering*, 2023 <https://doi.org/10.1016/j.eng.2023.04.013>
28. L. Wang, Heterogeneous data and big data analytics, *Autom. Control Inf. Sci.*, 2017, **3**(1), p 8-15.
29. C. Herriott and A.D. Spear, Predicting microstructure-dependent mechanical properties in additively manufactured metals with machine- and deep-learning methods, *Comput. Mater. Sci.*, 2020, **175**, p 109599.
30. K. Bobzin, W. Wietheger, H. Heinemann, S.R. Dokhanchi, M. Rom, and G. Visconti, Prediction of particle properties in plasma spraying based on machine learning, *J. Therm. Spray Technol.*, 2021 <https://doi.org/10.1007/s11666-021-01239-2>
31. R. Valente, A. Ostapenko, B.C. Sousa, J. Grubbs, C.J. Massar, D.L. Cote, and R. Neamtu, Classifying powder flowability for cold spray additive manufacturing using machine learning, in *2020 IEEE International Conference on Big Data (Big Data)* (IEEE, Atlanta, GA, USA, 2020), pp. 2919-2928. <https://doi.org/10.1109/BigData50022.2020.9377948>
32. D. Ikeuchi, A. Vargas-Uscategui, X. Wu, and P. King, Data-efficient neural network for track profile modelling in cold spray additive manufacturing, *Appl. Sci.*, 2021, **11**(4), p 1654.
33. D. Ikeuchi, A. Vargas-Uscategui, X. Wu, and P.C. King, Neural network modelling of track profile in cold spray additive manufacturing, *Materials*, 2019, **12**(17), p 2827.
34. M. Liu, H. Wu, Z. Yu, H. Liao, and S. Deng, Description and prediction of multi-layer profile in cold spray using artificial neural networks, *J. Therm. Spray Technol.*, 2021, **30**, p 1453-1463.
35. Z. Wang, S. Cai, W. Chen, R.A. Ali, and K. Jin, Analysis of critical velocity of cold spray based on machine learning method with feature selection, *J. Therm. Spray Technol.*, 2021, **30**(5), p 1213-1225.
36. C. Wang, X.P. Tan, S.B. Tor, and C.S. Lim, Machine learning in additive manufacturing: state-of-the-art and perspectives, *Addit. Manuf.*, 2020 <https://doi.org/10.1016/j.addma.2020.101538>
37. L. Breiman, Random forests, *Mach. Learn.*, 2001, **45**, p 5-32.
38. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)* (USA, 2016), pp. 785-794. <https://doi.org/10.1145/2939672.2939785>
39. L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, and A. Gulin, CatBoost: unbiased boosting with categorical features, in *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, USA, 2018), pp. 6639-6649
40. E.M. Senan, I. Abunadi, M.E. Jadhav, and S.M. Fati, Score and correlation coefficient-based feature selection for predicting heart failure diagnosis by using machine learning algorithms, *Comput. Math. Methods Med.*, 2021, **2021**, p 1-16.
41. A. Elhassan, S.M. Abu-Soud, F. Alghanim, and W. Salameh, ILA4: overcoming missing values in machine learning datasets an inductive learning approach, *J. King Saud. Univ. Comput. Inf. Sci.*, 2022, **34**(7), p 4284-4295.
42. D. Rubin, Inference and missing data, *Biometrika*, 1976, **63**(3), p 581-592.
43. L. Beretta and A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation, *BMC Med. Inform. Decis. Mak.*, 2016, **16**(3), p 74.
44. M.J. Azur, E.A. Stuart, C. Frangakis, and P.J. Leaf, Multiple imputation by chained equations: What is it and how does it work?, *Int. J. Methods Psychiatr. Res.*, 2011, **20**(1), p 40-49.
45. K.P.N.V. Satya Sree, J. Karthik, C. Niharika, P.V.V.S. Srinivas, N. Ravinder, and C. Prasad, Optimized conversion of categorical and numerical features in machine learning models, in *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (IEEE, Palladam, India, 2021), pp. 294-299. <https://doi.org/10.1109/I-SMAC52330.2021.9640967>
46. T. Chai and R.R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, *Geosci. Model Dev.*, 2014, **7**(3), p 1247-1250.
47. V.N.V. Munagala, R. Chakrabarty, J. Song, and R.R. Chromik, Effect of metal powder properties on the deposition characteristics of cold-sprayed Ti6Al4V-TiC coatings: an experimental and finite element study, *Surf. Interfaces*, 2021, **25**, p 101208.
48. M. Motz, J. Krauß, and R.H. Schmitt, Benchmarking of hyperparameter optimization techniques for machine learning applications in production, *Adv. Ind. Manuf. Eng.*, 2022, **5**, p 100099.
49. L. Shapley, in *A Value for n-Person Games*, ed. by H. Kuhn, A.W. Tucker, Contributions to the Theory of Games II (Princeton University Press, Princeton, 1953), pp. 307-317. <https://doi.org/10.1515/9781400881970>
50. S.M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, (Curran Associates Inc., Red Hook, NY, USA, 2017), pp. 4768-4777
51. R. Rodríguez-Pérez and J. Bajorath, Interpretation of machine learning models using shapley values: application to compound

- potency and multi-target activity predictions, *J. Comput. Aided Mol. Des.*, 2020, **34**(10), p 1013-1026.
52. M. Ahsan, M. Mahmud, P. Saha, K. Gupta, and Z. Siddique, Effect of data scaling methods on machine learning algorithms and model performance, *Technologies*, 2021, **9**(3), p 52.
 53. W. Ma, Y. Xie, C. Chen, H. Fukanuma, J. Wang, Z. Ren, and R. Huang, Microstructural and mechanical properties of high-performance Inconel 718 alloy by cold spraying, *J. Alloys Compd.*, 2019, **792**, p 456-467.
 54. T. Schmidt, F. Gärtner, H. Assadi, and H. Kreye, Development of a generalized parameter window for cold spray deposition, *Acta Mater.*, 2006, **54**(3), p 729-742.
 55. H. Assadi, F. Gärtner, T. Stoltenhoff, and H. Kreye, Bonding mechanism in cold gas spraying, *Acta Mater.*, 2003, **51**(15), p 4379-4394.
 56. A.S. Alhulaifi and G.A. Buck, A simplified approach for the determination of critical velocity for cold spray processes, *J. Therm. Spray Technol.*, 2014, **23**(8), p 1259-1269.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.