Journal of Materials Processing Tech. 315 (2023) 117908

Contents lists available at ScienceDirect

Journal of Materials Processing Tech.

journal homepage: www.elsevier.com/locate/jmatprotec

Tensile strength prediction in directed energy deposition through physics-informed machine learning and Shapley additive explanations

Clayton Cooper [a,1], Jianjing Zhang [a,2], Joshua Huang [a,3], Jennifer Bennett [b,c], Jian Cao [b,4], Robert X. Gao [a,*,5]

[a] Department of Mechanical & Aerospace Engineering, Case Western Reserve University, Cleveland, OH 44106, USA [b] Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA [c] Department of Civil and Mechanical Engineering, United States Military Academy, West Point, NY 10996, USA

ARTICLEINFO

Associate Editor: Dr. E. Tekkaya

Keywords:
Directed energy deposition
Tensile strength prediction
Interpretable machine learning
Shapley additive explanation
Model pruning

ABSTRACT

In directed energy deposition (DED), local material microstructure and tensile strength are determined by the thermal history experienced at each spatial location on the part. While prior research has investigated the effect of thermal history on mechanical properties, a tensile strength prediction model that is *physically interpretable* and *parsimonious* with good predictive *accuracy* is still needed. This paper investigates a data-driven predictive model with Shapley additive explanation (SHAP)-based model interpretation to address this issue. First, physically meaningful thermal features translated from prior experimental works are used as inputs to a neural network for tensile property prediction. SHAP values are then computed for the individual input features to quantify their respective influences on tensile property predictions and reduce model complexity using the metric of cumulative relative variance (CRV). Prediction of experimentally acquired Inconel 718 (IN718) tensile strength demonstrates that feature influences quantified by the developed method can be verified by findings from prior works, confirming the physical interpretability of the neural network prediction logic. In addition, model complexity reduction based on CRV has shown that fewer than 10% of the original features are required by the parsimonious model to achieve the same predictive accuracy of tensile strength as reported in prior literature, thereby demonstrating the effectiveness of SHAP-based feature reduction method in improving DED process characterization.

1. Introduction

Metal additive manufacturing (AM) has demonstrated its capability in producing parts with complex geometries, unique properties, and versatile applications in sectors including automotive, aeronautical, medical, and natural resource extraction (Vafadar et al., 2021). As a popular metal AM process, directed energy deposition, as illustrated in Fig. 1, has drawn broad and continuing interest (Dass and Moridi, 2019). Unlike powder bed processes, which are performed in an inert environment to prevent internal oxidation, DED shields the melt pool from oxidation using a stream of inert gas (Gibson et al., 2015). This process characteristic makes DED more flexible in terms of its material choice and build volume than powder bed methods.

To ensure DED product quality, issues such as microstructure, porosity, and residual stress need to be investigated. Notably, parts' local microstructure and mechanical properties are determined by the thermal history experienced at each on-part location during deposition. This history is characterized by repeated heating, cooling, and phase transformation, leading to an intricate process-property relationship and nonhomogeneous mechanical properties (Seifi et al., 2016). While thermal history prediction has been investigated, prediction of mechanical properties as a consequence of the thermal histories is of high interest to the AM community (Ness et al., 2021).

Characterizing the effect of thermal histories on microstructure (Hejripour et al., 2019) and directly predicting mechanical properties using thermal history-based methods is an emerging research topic (Snow et al., 2022). Complementing this is an ongoing influx of data-driven methods for mechanical property prediction in metal AM due to the flexibility and efficiency that these methods have demonstrated, as compared to model-based techniques (Qin et al., 2022).

Based on thermal history data, it has been observed that metals' local grain size, which is a determinant of mechanical properties, is inversely proportional

Received 20 June 2022; Received in revised form 16 November 2022; Accepted 8 February 2023
Available online 11 February 2023

* Corresponding author.
E-mail address: robert.gao@case.edu (R.X. Gao).
[1] 0000-0001-6212-8788
[2] 0000-0002-5760-6893
[3] 0000-0001-6397-8370
[4] 0000-0003-1023-5244 [5] 0000-0003-3595-3728
https://doi.org/10.1016/j.jmatprotec.2023.117908

to local temperature gradient and solidification boundary speed (Kou, 2003). High accuracy in tensile property prediction for Inconel 718 (IN718) has also been reported (Xie et al., 2021). Spatially localized temperature time series were extracted from in-situ infrared (IR) video of DED production and converted to time-frequency images through wavelet transform. A convolutional neural network (CNN) was then used to predict the material's tensile properties using wavelet scalograms as input to the network, with the relative error of the test data at about 6%. Random forest models (Breiman, 2001) were used to observe that the durations for which the temperature signal was within 654–857°C (cooling time) and 1260–1336 °C (solidification time) during in-machine cooling were good predictors of local tensile properties.

Similarly, a positive correlation between cooling time and tensile stresses and a negative correlation between solidification time and tensile stresses have been qualitatively observed (Glerum et al., 2021). The cooling and solidification times were also related to part microstructure, furthering the understanding of the DED process-structure-property relationship. Prior study (Bennett et al., 2021) confirmed the results of Glerum et al. by using cooling and solidification times to find quantitative correlations via an exponential equation.

To further improve the accuracy and robustness in DED tensile strength prediction, a model whose outcome is physically *interpretable* and *parsimonious* in its structure is desired. In the presented study, "interpretability" means that the influence of features can be analyzed to validate model behavior against physical findings from prior experimental studies, whereas parsimony means that model performance can be optimized while model complexity is simultaneously reduced (Salman and Liu, 2019).

Achieving physics-based interpretability enables users to assess a model for physical validity, synthesize new physical relationships, and improve understanding of model behavior (Guo et al., 2022). Specifically, if the prediction logic of a predictive model is spurious and deviates from the laws of physics, the model would learn physically insensible relationships and make incorrect predictions of material properties outside of the training dataset. Conversely, if model decision logic is consistent with known physics, new fundamental relationships may be discovered that merit further experimental verification for new knowledge generation (Sanchez et al., 2021).

Achieving model parsimony benefits the generation of a compact set of control variables that facilitates real-time DED process control. It has been shown that many different thermal characteristics may be used to predict mechanical properties, including time-frequency content (Xie et al., 2021) and time spent in certain temperature regions (Bennett et al., 2021). Further clarification on how these features compare to one another in terms of predictive power would lead to better identification of the most relevant predictors and a more *compact* set of influential thermal variables in DED. Finally, achieving high accuracy is essential for predicting part quality and system throughput in DED, as the above-described facets of interpretability cannot come at the cost of substantial degradation of model accuracy.

The presented study aims to develop a DED part property prediction model that meets these criteria. For this purpose, a neural network has been developed to predict the ultimate tensile stress (UTS) and yield stress (YS), collectively referred to as "tensile strength," of thin-walled IN718 DED components, using thermal history-based features, DED process physics, and findings from prior experimental studies as the network inputs. In addition, Shapley additive explanation (SHAP) values are investigated to quantify the influence of each of the model input features and assess the model's physical interpretability and validity (Lundberg and Lee, 2017). Finally, the method of cumulative relative variance (CRV) is investigated to efficiently remove less influential input features based on their SHAP-quantified influence to achieve model parsimony.

This study's outcome is a prediction model that has six orders of magnitude fewer parameters than the best model currently reported in the literature, while achieving comparable or higher predictive accuracy. An overview of the developed method is shown in Fig. 2, with the interconnection among the key components of the method as well as the associated details of the key components summarized.

The rest of this paper is organized as follows: Section 2 presents the background and procedure of the developed method including an introduction to SHAP values and their utility in model interpretation and feature pruning. Section 3 describes a DED experiment to evaluate the method, using IN718 as the test material. The result of model prediction, interpretation, and complexity reduction (i.e., feature pruning) is discussed in Section 4. In Section 5, conclusions and directions for future work are provided.

2. Theoretical background

Given an initial set of thermal history-derived tensile strength predictors, i.e., *input features*, three steps are needed for developing an interpretable tensile strength prediction model: 1) regression modelling, 2) feature influence quantification, and 3) model complexity reduction.
Each of these steps is formulated in the following subsections.

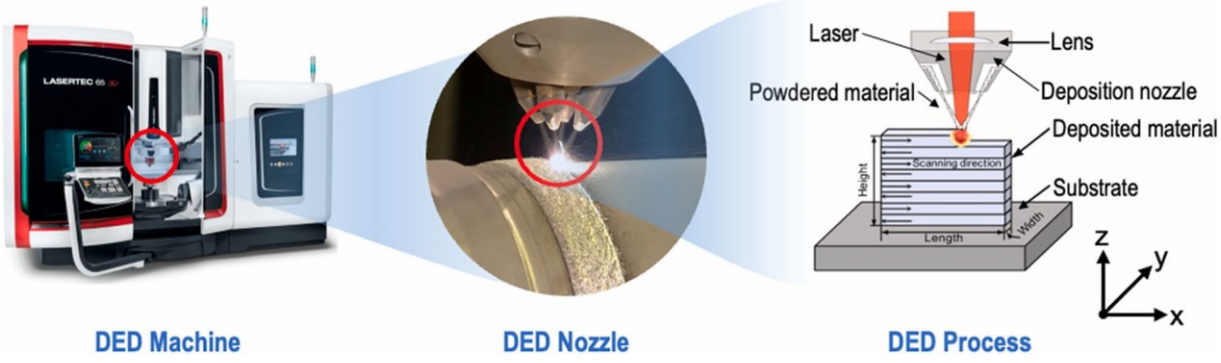

Fig. 1. DED process diagram; part width in *y* direction is equivalent to one laser diameter, making the part "thin-walled".
adapted from DMG Mori, Optomec, and (Zhang et al., 2020)

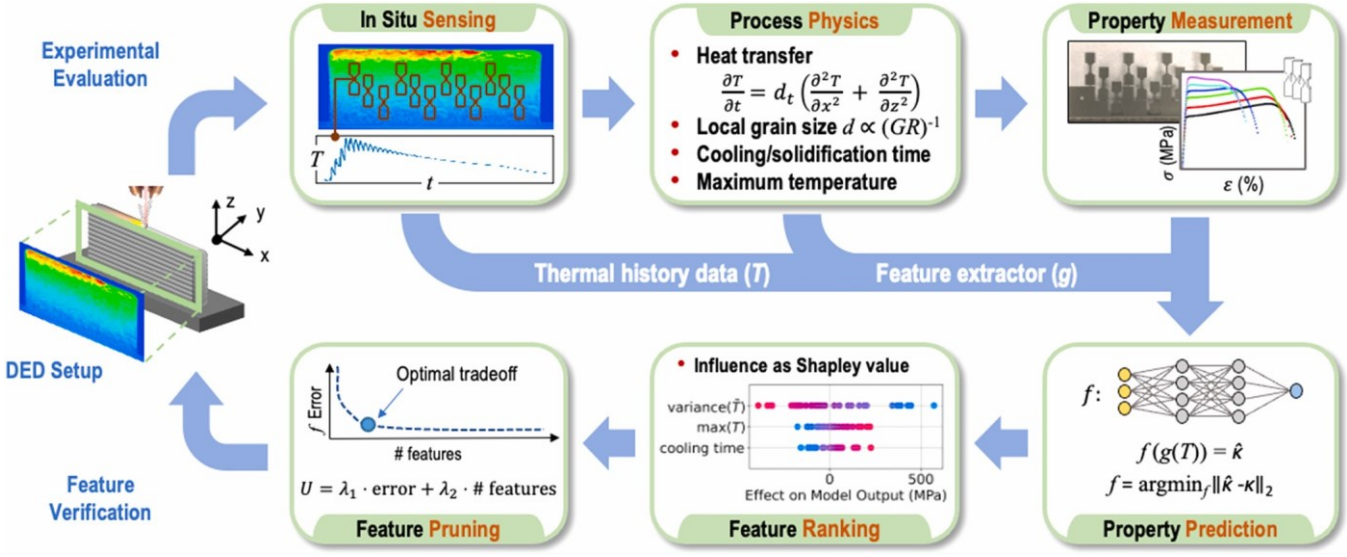

Fig. 2. Interpretable tensile strength prediction for DED using machine learning and Shapley additive explanations.

2.1. Regression modeling

The objective of tensile strength prediction is to establish a regression model that predicts tensile strength with minimized deviation from experimentally measured values that serve as the ground truth. Mathematically, such a tensile strength prediction model f can be formulated as:

$$f(g[\boldsymbol{T}_i(t,x,z)],\mathbf{w}) = \hat{\kappa}_i \mid \mathbf{w} = \underset{\mathbf{w}}{\text{argmin}} \left[\frac{1}{N} \sum_{i=1}^{N} \|\kappa_i - \kappa_i\|_2 \right] \quad (1)$$

where g is a feature extractor, $\boldsymbol{T}_i(t, x, z)$ is the i^{th} time-varying and spatially localized temperature field acquired in the coordinate frame of Fig. 1 (see the "DED Process" block), $\mathbf{w}$ represents the optimizable parameters of f, $\hat{\kappa}_i$ and κ_i represent the predicted and measured tensile strength at location (x, z) corresponding to $\boldsymbol{T}_i$, respectively, and N is the total number of measured tensile strength values in the experiment. Model f is therefore data-driven, since its $\mathbf{w}$ parameters are optimized based on training tuples $(\boldsymbol{T}_i, \kappa_i)$, and able to analyze thermal history features in DED, as extracted by g.

To ensure that f's prediction logic is consistent with physics to achieve physics-based model interpretability, the influence of each input feature in determining $\hat{\kappa}_i$ needs to be assessed. This requires that known relationships between input variables and tensile strength predictions be maintained, and that a change in a feature's value has a physically consistent effect on the model output. Examples include the direct proportionality between cooling time and YS reported in (Bennett et al., 2021) and (Glerum et al., 2021), as well as the proportionalities between temperature ranges and UTS reported in (Xie et al., 2021). Violation of these relationships indicates that f has not learned to obey the known physical mechanisms governing DED. Further discussion of this aspect of interpretability is provided theoretically in Section 2.2 and experimentally in Section 4.2.

Model parsimony is another reflection of model interpretability. A more parsimonious model will have fewer input features. Accordingly, fewer interactive effects among features need to be accounted for. In the presented study, model parsimony is achieved by the joint minimization of model error and input size as shown below:

$$f_{\text{opt}}, g_{\text{opt}} = \text{argmin}_{f,g} \left[\lambda_1 \cdot \underbrace{\frac{1}{N} \sum_{i=1}^{N} \|\kappa_i - \kappa_i\|_2}_{\text{model error}} + \lambda_2 \cdot \underbrace{\text{features}(g)}_{\text{input size}} \right] \quad (2)$$

where f_{opt} and g_{opt} are optimal f and g, λ_1 and λ_2 are weighting coefficients, and features(g) is the number of features extracted from thermal history data. The f and g resulting from (2) are, by definition, optimally parsimonious.

2.2. SHAP-based model interpretation

To understand model decision logic and achieve interpretability, the *influence* of each input feature on the outcome of model predictions needs to be quantified. This can be achieved by investigating the SHAP values corresponding to each input feature, and their variances.

The concept of the SHAP algorithm (Lundberg and Lee, 2017) originates from game theoretic problems in economics, which determine how to fairly distribute "payout" for a team of players based on their respective contribution to the task outcome (Shapley, 1951). By analogy, the SHAP algorithm in the context of machine learning (ML) in general is to determine the "importance" of each of the ML model's input variables based on their respective influence in affecting the model's output. Knowing the "importance" of input variables allows for the interpretation and verification of the model predictive logic with empirical or physical knowledge, contributing to the acceptance of ML as a complement to physics-based reasoning by human experts (Roscher et al., 2020).

In the specific context of additive manufacturing, the input variables of interest are the features from in-situ IR sensing data (such as temperature variance and maximum temperature) that are envisioned to influence the model's output (such as part ultimate tensile strength). Quantifying the "importance" of these input variables using SHAP allows for identifying the true influential variables, which then serves as the basis for process control and optimization, as well as for simplifying the predictive model by removing the "insignificant" ones.

Define predictive model f, input vector $\mathbf{x}_i$, and feature j such that $\mathbf{x}_i[j]$ represents the j^{th} feature of the i^{th} input sample. The influence of feature j on $f(\mathbf{x}_i)$ can then be interpreted as the *change* in $f(\mathbf{x}_i)$ caused by j. How specifically a feature j would change a model's output depends on f's learned decision logic and the nonlinear weighting, $\mathbf{w}(\mathbf{x}_i[j])$ that is learned for each feature.

To observe feature influence, consider a simplified situation where features are incrementally considered by a model, as shown in Fig. 3. Before any feature is considered, the model prediction is the unconditioned expectation of the model output, $\mathrm{E}(f)$. With only the 1st feature known, the model takes its best guess with the information it has available at the time, i.e., the expected output *conditioned* on the 1st feature, $\mathrm{E}(f \mid j_1)$. The influence of j_1 can be quantified as the model change as a consequence of j_1, $\mathrm{E}(f \mid j_1) - \mathrm{E}(f)$. This is true for subsequent features as well, e.g., the influence of j_2 is $\mathrm{E}(f \mid j_1, j_2) - \mathrm{E}(f \mid j_1)$. Fig. 3 also shows that low-influence terms, such as j_5, will contribute little "correction" to the model's output. If j_5 was removed, the model prediction would deviate slightly from the full feature model, but the model would be simpler, have fewer free parameters, be less prone to overfitting, and computationally more efficient.

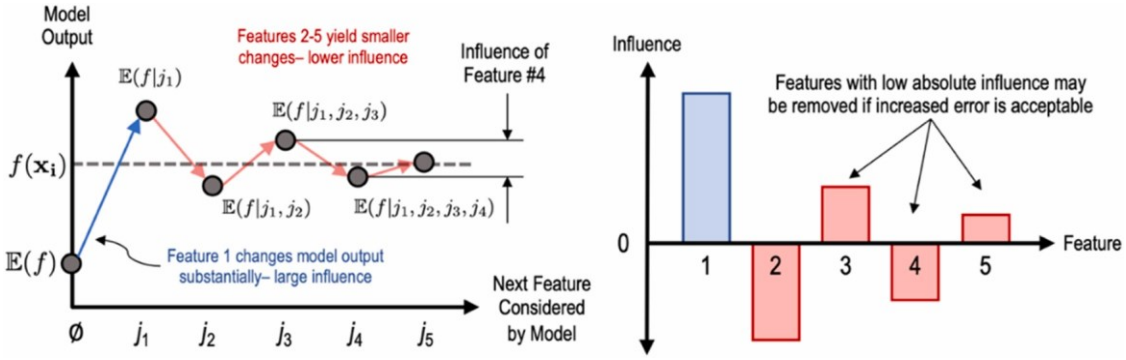

Fig. 3. Example of model output change as a measure of feature influence.

To systematically determine the influences for every feature j and sample i, SHAP values are investigated. Assume a candidate input feature set F of size $|F|$, feature index $j \in [1, |F|]$, feature subsets $S \subseteq F \setminus \{j\}$ and $S^* = S \cup \{j\}$, and predictive model $f : \mathbb{R}^{\beta} \rightarrow \mathbb{R}$, where β is the size of the input to f that can vary depending on the number of features being considered. The SHAP value for feature j of sample i ($\boldsymbol{\phi}_i[j]$) determines j's influence by finding the *average* change in model response caused by j, evaluated across all possible S to account for feature interaction effects.

Mathematically, $\boldsymbol{\phi}_i[j]$ is defined as:

$$\phi_i[j] = \sum_{S,\,|S|,\,|F|\,|F|-|S|-1} \left[f_{S^*}(\mathbf{x}_i[S^*]) - f_S(\mathbf{x}_i[S]) \right] \tag{3}$$

where $f_{S^*}(\mathbf{x}_i[S^*]) - f_S(\mathbf{x}_i[S])$ is the change in model output induced by feature j (Young, 1985). Accordingly, $\boldsymbol{\phi}_i[j]$ is the weighted average influence of feature j on $f(\mathbf{x}_i)$ across all S, with weights given by the inverse multinomial coefficient in (3). SHAP values are found for a whole dataset by calculating (3) for all samples and feature combinations.

Once SHAP values are computed for all samples and features using (3), model prediction $f(\mathbf{x}_i)$ can be *decomposed* in the additive form (Shapley, 1951) as shown in (4):

$$\sum_j f(\mathbf{x}_i) = \mathrm{E}(f) + \phi_i[j] \tag{4}$$

SHAP values for influential features will constitute most of this sum whereas weak features will contribute little to $f(\mathbf{x}_i)$. As a result, feature importance for $f(\mathbf{x}_i)$ can be quantified, model prediction logic can be interpreted, and noninfluential features can be removed. In addition, (3) holds for any f (Lundberg and Lee, 2017). As a result, SHAP values can transform the feature influence of any predictive model into the linear additive form as shown (4) to facilitate interpretability even if f is nonlinear, such as a neural network. However, direct computation of (3) requires training a different f for each possible S. This means the computational complexity of finding each $\boldsymbol{\phi}_i[j]$ scales according to $2^{|F|}$, which rapidly becomes impractical. To overcome this, the Kernel SHAP method has been introduced for SHAP value approximation (Lundberg and Lee, 2017). This technique trains f once and then linearly approximates $f(\mathbf{x}_i)$ at each $\mathbf{x}_i$ to satisfy (4). Each linear approximation is found by repeatedly perturbing $\mathbf{x}_i$ and fitting a linear model to the perturbed ($\mathbf{x}_i$, $f(\mathbf{x}_i)$) pairs. The coefficients of the fitted linear model approximate true SHAP values for $\mathbf{x}_i$. The computational complexity of Kernel SHAP is $O(i)$ rather than $O(2^{|F|})$ as in (3), making it computationally practical.

2.3. Feature pruning and model complexity reduction

After acquiring SHAP values, the aggregate influence of feature j on f can be quantified as the variance of j's SHAP values across all input samples, represented as $\boldsymbol{\sigma}^2[j]$. If $\boldsymbol{\sigma}^2[j] \rightarrow 0$, feature j has the same influence on f no matter the value of $\mathbf{x}_i[j]$, so the feature influence can be thought of as a constant. In many regression models, including neural networks, incorporating such a constant is equivalent to learning a bias term. Therefore, if feature j is removed and f is retrained, a bias term can be learned in place of the removed feature.

To rank the features' influence, SHAP value *relative variance* (RV) is calculated, which is defined as:

$$\mathrm{RV}[j] = \boldsymbol{\sigma}^2[j] \Big/ \sum_{j \in F} \boldsymbol{\sigma}^2[j] \tag{5}$$

where the numerator is the j^{th} feature's SHAP value variance and the denominator is the sum of SHAP value variances of every feature in the full feature set. Features with the lowest SHAP RV are removed from the feature set and converted to bias terms because they contribute the least variety to the model output and are therefore well-suited to become learned biases instead of inputs. Because of this ordered feature removal rule (always remove the feature with the lowest RV), the number of feasible feature subsets is reduced from $2^{|F|}$ to $|F|$.

While using RV rankings reduces the number of feature subsets to $|F|$, searching through these subsets may still be impractical if $|F|$ is large or if f is time-consuming to optimize. For example, if $|F| = 20$ and f requires 5 min to train, then exhaustive feature subset exploration would take 20 feature subsets × 5 min/subset = 100 min. To expedite the feature reduction process, a coarse search strategy is developed by considering the subsets' ratio of SHAP value variance to the size of the subset.

Optimal feature subsets (S_{opt}) should contain features that explain as much of the SHAP value variance as possible (maximizing predictive power) while also minimizing the number of features needed to achieve that explanation (maximizing parsimony). Considering that two subsets with similar explained variance (e.g., 90% vs. 89%) should theoretically have similar prediction performance, the amount of explained variance should be *coarsely* varied during the search over the $|F|$ remaining feature sets. Specifically, finding S_{opt} for $K \ll |F|$ coarsely spaced levels of explained variance will give K feature subsets with varying performance-parsimony tradeoffs while further reducing the feature subset search space size from $|F|$ to K.

The optimal feature subset S_{opt} for explained SHAP variance percentage ℓ is defined as:

$$S_{opt,\,\ell} = \underset{S}{\operatorname{argmax}} \frac{\mathrm{CRV}}{|S|}(S) \text{ s.t. } \mathrm{CRV}(S) \geq \ell$$

$$\mathrm{CRV}(S) = \sum \mathrm{RV}[j] \tag{6}$$

$j \in S$

where *cumulative relative variance* (CRV) ∈ [0,1] is the sum of RV values for all features in feature subset S. By definition, CRV(S) is the proportion of SHAP value variance explained by S. Because of the argmax operation in (6), S_{opt} must contain the $|S|$ features with the largest RV values and no others. In the presented study, a total of K CRV values that are logarithmically spaced are chosen to expedite the search process through the $|F|$ feature subsets remaining after the RV ranking process. Logarithmic spacing is chosen based on experimentally observed CRV curves. These K feature subsets can then be compared on the bases of parsimony and performance.

There are several alternative methods to CRV that can quantify the "importance" of a feature in predictive model:

1) Mean SHAP value is chosen as the importance metric in (Marcilio and Eler, 2020), namely, features with higher mean SHAP values are considered more "influential" than the ones with lower mean SHAP values. However, in our predictive model, the mean SHAP value of each input features is shown to hover around zero, indicating a bidirectional influence from each of these features on the part property. As a result, the overall "magnitude" of the feature's influence cannot be reflected by its mean value. In contrast, the variance as reflected in the CRV of SHAP values shows such "magnitude" of influence, as larger variance indicates more change in the part property can be induced by the change in the corresponding feature as compared to the feature with smaller variance.
2) Maximum SHAP value is used to assess feature importance in (Mokdad et al., 2015), with a higher maximum indicating a more important feature. While this approach indicates the upper bound of each feature's influence, it is highly sensitive to outlying SHAP values and neglects to consider the expected influence of each feature.
3) Constructing "neighborhoods" of candidate feature sets based on their SHAP values is proposed in (Chen et al., 2021). In this method, p random feature subsets are first used to train p predictive models. The model with the highest average SHAP value across all features has a "neighborhood" of feature sets built around it by adding or removing a feature. However, determination of the best candidate required training and testing of separate models defined for every feature set in the neighborhood, which can be computationally impractical. In contrast, CRV can be directly obtained from raw SHAP values, and does not require additional model training and testing.

Based on the above discussion, CRV is chosen as the metric for determining the optimal feature set in this study.

It is noted that bidirectional step-wise regression techniques (Hocking, 1976) have been traditionally used for feature selection in predictive models, where exactly 1 feature is added or removed at a time. Compared to SHAP-based selection, step-wise techniques have the following limitations:

1) **Computational efficiency**. At each iteration, the bidirectional step- wise regression is required to evaluate the model and quantify the contribution for each candidate input variable individually to determine which variable is added to or removed from the model for the next iteration. As the number of features increases linearly, the number of input variables can increase exponentially when interactions among multiple features are considered. Since the feature screening process generally involves a large set of candidate features at the beginning, using bidirectional step-wise regression is expected to result in significant computational cost (Zoga ła-Siudem and Jaroszewicz, 2021).
2) **Global optimum**. As only one input variable can be added or removed at each iteration, the bidirectional step-wise regression is a "greedy" algorithm by nature (i.e., the optimality is determined only by the immediate, local improvement resulted from the addition or removal of an input variable, rather than globally by accounting for all remaining iterations down the road). As a result, the algorithm may characterize only a small subset of all combinations of input variables and miss the global optimal combination.

In contrast, SHAP implicitly and *simultaneously* characterizes *all* possible feature interactions to determine globally the "importance" of each feature as SHAP value variance, which only requires the model to be evaluated once.

2.4. Comparing CRV-generated feature sets

After generating K feature subsets at K logarithmically spaced CRV levels as described above, an additive utility function (Malakooti, 2014) is investigated to find which subset of the features best balances parsimony and predictive accuracy. The best candidate is the solution to the motivating simultaneous optimization problem as described in (2). Additive utility functions take the form shown in (7):

$$U_k = \langle \boldsymbol{\eta}, \mathbf{m}_k \rangle \quad \boldsymbol{\eta} \in [0,1]^P, \sum \boldsymbol{\eta} = 1, \mathbf{m}'_k \in [0,1] \quad |_P \tag{7}$$

where $k \in [1, K]$ is the index of the CRV-generated feature subset being considered, U is a measure of utility to be maximized, $\boldsymbol{\eta}$ is a P-length vector of weights, and $\mathbf{m}'_k$ is a P-length vector of standardized performance metrics for feature subset k. Eq. (7) allows for k different prediction models to be compared on the bases of P different performance metrics simultaneously; the model with the highest U value is the best. Additionally, weight vector $\boldsymbol{\eta}$ allows each metric to be assigned its own weight or "importance" in the utility calculation. This approach therefore gives a systematic way of comparing models while considering disparate performance criteria in tandem. Such a system is necessary to solve the optimization problem posed in (2).

To fairly compare the P metrics, they must be standardized to the same scale. This is to prevent metrics with large magnitudes of (10^1–10^3, e.g., absolute error) from dominating metrics with small magnitudes (10^{-3}–10^0, e.g., percent error) in the calculation of each U. Standardized metric vector $\mathbf{m}'_n$ is found from unstandardized performance metric vector $\mathbf{m}_n$:

$$\mathbf{m}'_n[p] = \begin{cases} \mathbf{m}_n[p] - \min_P(\mathbf{m}_n[p]) \text{ if } \mathbf{m}_n[p] \text{ to be maximized} \\ \max_P(\mathbf{m}_n[p]) - \mathbf{m}_n[p] \text{ if } \mathbf{m}_n[p] \text{ to be minimized} \end{cases} \Big/ \begin{cases} \max_P(\mathbf{m}_n[p]) - \min_P(\mathbf{m}_n[p]) \text{ if } \mathbf{m}_n[p] \text{ to be maximized} \\ \max_P(-\mathbf{m}_n[p]) - \min_P(-\mathbf{m}_n[p]) \text{ if } \mathbf{m}_n[p] \text{ to be minimized} \end{cases} \tag{8}$$

where $p \in [1, P]$ is the index of the metric being normalized, $\mathbf{m}_n[p]$ is the p^{th} element of $\mathbf{m}_n$, $\min_P(\cdot)$ is the minimum of performance metric values across all P, and $\max_P(\cdot)$ is the maximum of performance metric values across all P. This standardization process ensures that all metrics are on the same [0,1] scale and can be fairly compared to one another regardless of unscaled metric magnitude.

The need for maximization or minimization via (8) is dependent on the metric being considered. For instance, error metrics such as mean absolute error should be *minimized* since lower error indicates more accurate predictions. On the other hand, metrics such as correlation coefficient should be *maximized* since higher correlation indicates more accurate predictions.

Finally, the elements of $\boldsymbol{\eta}$ are application dependent. Different applications have different priorities, e.g., interpretability, accuracy, computation time, and these priorities are reflected in the weights and the selection of performance metrics used in $\mathbf{m}'_n$. The prediction model and feature subset which maximize U given $\boldsymbol{\eta}$ and $\mathbf{m}'_n$ are the f and g that satisfy (2).

In summary, SHAP values from Kernel SHAP, feature rankings from CRV, and the utility function of (8) work in tandem to solve (2) and identify the tensile strength prediction model and corresponding feature subset with the best

tradeoff between performance and interpretability. **3. Experiment, feature extraction, and model fitting**

This section presents the DED part production and data acquisition experiment that involves tensile testing, feature extraction, and model fitting.

3.1. DED and tensile testing

Twelve thin walls of IN718 were deposited on a substrate of SS304 using a DMG MORI LaserTec 65 3D DED machine, as illustrated in the process schematic of Fig. 1. All walls were 60 mm tall. Three of the walls were 80 mm long whereas the other nine were 120 mm long. The powdered material's granule diameters were between 50 and 150 μm and were delivered to the laser-generated melt pool at 18 g/min using an argon gas stream with 7 L/min flow. The 1800 W laser's spot diameter was 3 mm. It traversed the walls at 17 mm/s.

Three deposition techniques were investigated: 1) continuous, 2) dwell, and 3) melt pool control. In continuous deposition, the laser traversed the length of the walls with no pauses and no changes to the deposition parameters. The dwell technique deactivated the laser and powder delivery systems for 5 s after each layer before depositing the next one. Melt pool control modulated the laser power during deposition to keep the melt pool size constant. A coaxial CCD camera with a resolution of 180 × 250 pixels and pixel size 30 μm^2 was used to view the melt pool and count the number of pixels with intensity greater than a user-defined threshold. A PID controller then modulated the laser power to keep the number of pixels above the threshold, representing the melt pool size, between user-specified bounds (Bennett et al., 2018).

The time-varying temperature field of each wall ($\boldsymbol{T_i}$) was monitored perpendicular to the scanning direction using a FLIR A655sc IR camera, as seen in Fig. 4. The camera's 640 × 480 pixel resolution captured a 128 × 96 mm field of view, resulting in a pixel size of 200 × 200 μm. Prior to the experiment, the recorded IR temperatures were calibrated to match the real temperatures using the methodology detailed in (Bennett et al., 2021). Essentially, the measured IR values were corrected by observing the liquidus-solidus transition point in the IR data using the data's 2nd derivative and setting this transition point equal to the known liquidus-solidus transition temperature of IN718. Following deposition, 135 miniaturized tensile coupons were cut from the walls using electrical discharge machining (EDM). The coupons were then tested on a Sintech 20 G tensile test machine at a strain rate of 0.02 mm/s until fracture occurred, and their UTS and YS values were recorded. The samples were tested as-built, no heat treatment was applied. Thermal features, as described in Section 3.2 and shown on the right side of Fig. 4, were calculated from the IR video at the center of each tensile coupon. These were then processed to become predictive model inputs as described in the following section.

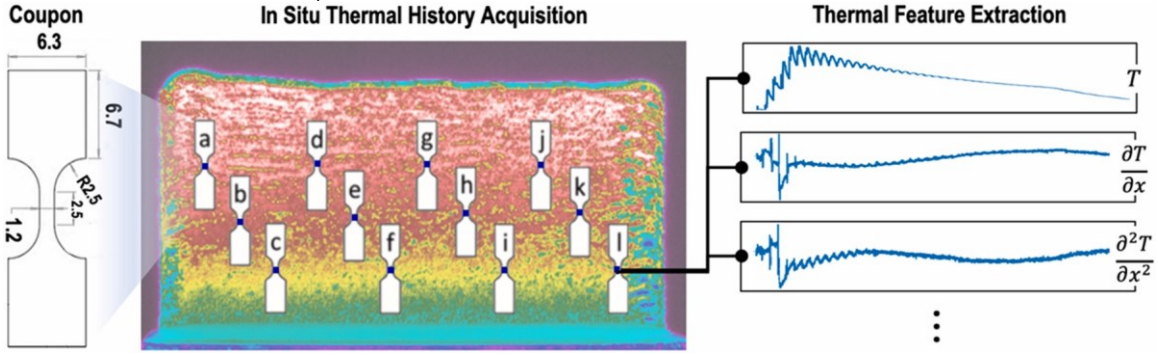

Fig. 4. Thermal history monitoring and data dimensions, adapted from (Glerum et al., 2021).

3.2. Feature selection

To extract information embedded in the DED temperature fields measured by the IR camera for tensile strength prediction, a set of predictors is first formulated to generate a comprehensive feature set based on knowledge gained in previous work. Subsequently, the feature set is pruned using the SHAP value relative variance method as discussed before. The outcome of the process is that only the predictors of high relevance to tensile strength are retained. The process also enables a direct comparison of the influence of the IN718 tensile strength predictors as studied in previous experiments.

To determine the initial feature set from *T,* physical validity and DED process controllability are considered. Notably, thermal evolution in DED can be characterized by extracting the 1) first and 2) second spatial derivatives of the temperature field of each wall with respect to the *x*- direction, *z*-direction, and time (*t*). These partial derivatives are represented as the set $\Delta(T) = \{\partial T/\partial x, \partial T/\partial z, \partial T/\partial t, \partial^2 T/\partial x^2, \partial^2 T/\partial z^2, \partial^2 T/\partial t^2\}$. Derivatives in the *y*-direction are not considered due to the symmetric heat flux about the thin wall's *y*-axis as seen in Fig. 1. Since the elements of Δ(*T*) determine the material properties of the DED parts after production (Bennett et al., 2018), they are expected to provide both predictive power and physical interpretability of the predictive model developed in this study.

Since the tensile testing provided the mechanical properties at the center of each coupon, 10 pixel × 10 pixel regions of interest (ROIs) at each coupon center were established as shown by the blue dots in Fig. 4. Average signals for *T* and Δ(*T*) were then found for each ROI. This was done to reduce error in temperature measurement caused by emissivity changes due to roughness of the surface parallel to the camera (Bennett et al., 2018).

3.2.1. First partial derivatives

First partial derivatives ($\partial T/\partial x$, $\partial T/\partial z$, and $\partial T/\partial t$) are considered because they are inherently related to local grain size (*d*) and shape, which in turn affect the tensile strength. As the grain size *d* decreases, a material experiences more internal friction and tensile strength increases (Callister and Rethwisch, 2012). In (9), the relationship between *d* and the first partial derivatives is summarized (Kou, 2003):

$$d \propto (GR)^{-1} \quad \propto \frac{\partial T}{\partial x}, \frac{\partial T}{\partial z}; R \propto \frac{\partial T}{\partial t} \therefore d \propto \left(\frac{\partial T}{\partial x}, \frac{\partial T}{\partial z}, \frac{\partial T}{\partial t}\right)^{-1} \quad G\ \partial x\ \partial z\ \partial t\ \partial x\ \partial z\ \partial t \quad (9)$$

where *G* is the spatial temperature gradient in the *x-z* plane (°C/m) and *R* is the cooling rate (°C/s). Additionally, grain shape is determined by the ratio of *G/R*. A high *G/R* induces planar grains whereas low *G/R* induces equiaxed dendritic grains (Kou, 2003). This morphology subsequently determines tensile properties in that higher planarity is associated with higher ductility and reduced UTS and YS (Gockel et al., 2014).

3.2.2. Second partial derivatives The second derivatives $\partial^2T/\partial x^2$ and $\partial^2T/\partial z^2$ are considered because they represent local material density (ρ), as reflected in the heat equation (Moran, 2003) as expressed in (10):

$$\frac{\partial T}{\partial t} = \Omega \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial z^2} \quad \overline{x^2} \quad \overline{z^2} \quad \overline{t} \quad \Omega \propto \rho^{-1} \therefore \left(\frac{\partial^2 T}{\partial} + \frac{\partial^2 T}{\partial}\right) \Big/ \left(\frac{\partial T}{\partial}\right) \propto \rho \tag{10}$$

Interstitial defects which reduce tensile strength, such as pores and lack of fusion, will manifest themselves as lower local density, which is detectable using the derivative-based approach of (10). Given that material density is positively correlated with UTS and YS (Sun et al., 2016), $\partial^2T/\partial x^2$ and $\partial^2T/\partial z^2$ are expected to be positively correlated with UTS and YS as well.

Finally, the second time derivative $\partial^2T/\partial t^2$ is also calculated since it is an indicator of the liquidus-solidus transition in DED (Bennett et al., 2018), which also affects the tensile strength.

3.2.3. Tabular representation of time series signals

The above-described feature selection process yields a total of seven model input features for each coupon: T and the six elements of $\Delta(T)$ given by $\{\partial T/\partial x, \partial T/\partial z, \partial T/\partial t, \partial^2T/\partial x^2, \partial^2T/\partial z^2, \partial^2T/\partial t^2\}$. However, each of these features are time-varying signals. Considering that signals with long duration pose a challenge to recurrent prediction models and that incorporating temporal attention into a predictive model (Bai et al., 2018) to avoid forgetting information would rapidly increase the number of model parameters and lead to overfitting, the seven model input features are characterized using low-dimensional statistical representations such as mean and variance. This simplifies feature influence interpretation since the time dimension is eliminated and thus influence will not vary as a function of time.

Five statistical measures are calculated for T and $\Delta(T)$: maximum, mean, variance, skewness, and kurtosis (Everitt and Skrondal, 2010). Prior study has found that maximum temperature is a determinant of subsurface porosity in metal AM (Paulson et al., 2020). Therefore, maxima are calculated for T and the elements of $\Delta(T)$ and included in the prediction model to achieve the same predictive power as seen in literature. As for the other statistics, since skewness is used to scalarize time-varying signals, it can be interpreted as a measure of *time* spent away from the signal mean. Positive skewness indicates that a signal's value is predominantly above the mean, and vice-versa for a negative skewness. Similarly, kurtosis represents the *time* spent at extreme values. Low kurtosis indicates that the signal consistently remained near its mean value whereas high kurtosis indicates more time around extremes.

The moment statistics are calculated during the final cooling phase (T) as shown in Fig. 5, considering that this phase predominantly determines the final part microstructure (Kou, 2003). Microstructures formed prior to this phase become irrelevant because they are "reset" by reliquification of the ROI, whereas temperatures after this phase are not high enough to induce microstructural changes (Bennett et al., 2021). This means that the temperature signal outside of this phase does not have a lasting effect on tensile strength and should not be considered by the mean, variance, skewness, and kurtosis. $T$is defined by the liquidus temperature, T_L, and lower critical temperature, T_A (Clarke, 2014), as shown in Fig. 4. For IN718, T_L = 1364 °C and T_A = 654 °C (Bennett et al., 2021).

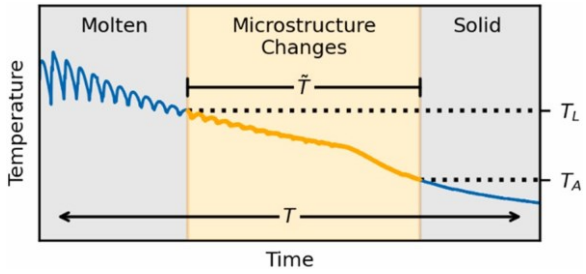

Fig. 5. Temperature signal truncation ($T \rightarrow \tilde{T}$) during final cooling phase.

3.2.4. Additional features

In addition to the 35 statistical features described above (7 signals × 5 statistics), the cooling and solidification times as defined in (Xie et al., 2021) are found for T to comprehensively compare tensile strength predictors proposed in (Kou, 2003) and (Farshidianfar et al., 2016). These works elucidated the significance of the cooling and solidification times in determining part microstructure and mechanical properties in AM, namely that fast solidification results in material with higher strength while more cooling time in the specified temperature range (Xie et al., 2021) strengthens the material by allowing more time for δ-phases to merge and break the brittle Laves phase.

Finally, to encode the three deposition control strategies, i.e., continuous, dwell, and melt pool control, a three-element one-hot vector is used. If temperature field $\boldsymbol{T}_i$ is a result of deposition type $v \in \{1, 2, 3\}$, the v^{th} element of the one-hot vector is set to 1 while all other elements are set to 0. This vector, when appended to the 35 statistical measures for T and $\Delta(T)$ along with cooling and solidification times, results in 40 tabular features to be used as inputs to tensile strength predictive model. These inputs are shown in Table 1.

3.3. Model training

A multilayer perceptron (MLP) was investigated as the tensile strength prediction model, f, since it is generally suited for approximating non-affine and continuous functions with arbitrary precision (Cybenko, 1989). Optimizing parameters $\mathbf{w}$ of f, as defined in (1), is achieved via backpropagation.

Two MLPs were trained using all 40 tabular features: one for UTS and one for YS, respectively. The 135 tensile coupons were split into training and testing sets (95 and 40 coupons, respectively). These datasets were stratified based on the spatial location of the tensile coupons (*a*–*l*) specified in Fig. 4, as well as the three deposition types. Each model was trained until testing loss did not decrease for 500 consecutive epochs, at which point the MLP weight matrices associated with the lowest recorded testing loss were restored.

MLP hyperparameters which yielded low prediction error were obtained via search over the following hyperparameter ranges: 1) 4–8 hidden layers, 2) 2–5 nodes per layer, 3) leaky ReLU (Maas et al., 2013) and swish (Ramachandran et al., 2017) activation functions, 4) $||\cdot||_1$ and $||\cdot||_2$ weight regularizers, 5) ω = 0.3–0.7, 6) 1×10^{-5}–1×10^{-3} initial learning rate, 7) 50,000–200,000 decay steps, 8) Adamax (Kingma and Ba, 2017) and Nadam (Dozat, 2016) optimizers, and 9) batch sizes of 4, 8, and 16, respectively.

The number of hidden layers and nodes/layer were bounded based on trial-and-error experimental observations: models smaller than 4 layers × 2 nodes/layer failed to learn mappings from the thermal features to tensile properties and models larger than 8 layers × 5 nodes/ layer consistently overfit the training data. A similar process was used to set the bounds for the weight regularization factor (ω), initial learning rate, and the number of learning rate decay steps. Learning was sluggish if ω was greater than 0.7, the initial learning rate was less than 1×10^{-5}, or the number of decay steps was below 50,000. On the opposite end of the spectrum, learning was unstable and tended towards exploding gradients if ω was less than 0.3, the initial learning rate was greater than 1×10^{-3}, or the number of decay steps was greater than 200,000.

Table 1
Inputs to tensile strength prediction model.

Partial Derivatives	mean(T)	$\max(T)$	variance(T)	skewness(T)	kurtosis(T)
	$\max(\partial T/\partial x)$	mean($\partial T/\partial x$)	variance($\partial T/\partial x$)	skewness($\partial T/\partial x$)	kurtosis($\partial T/\partial x$)
	$\max(\partial T/\partial z)$	mean($\partial T/\partial z$)	variance($\partial T/\partial z$)	skewness($\partial T/\partial z$)	kurtosis($\partial T/\partial z$)
	$\max(\partial T/\partial t)$	mean($\partial T/\partial t$)	variance($\partial T/\partial t$)	skewness($\partial T/\partial t$)	kurtosis($\partial T/\partial t$)
	$\max(\partial^2 T/\partial x^2)$	mean($\partial^2 T/\partial x^2$)	variance($\partial^2 T/\partial x^2$)	skewness($\partial^2 T/\partial x^2$)	kurtosis($\partial^2 T/\partial x^2$)
	$\max(\partial^2 T/\partial z^2)$	mean($\partial^2 T/\partial z^2$)	variance($\partial^2 T/\partial z^2$)	skewness($\partial^2 T/\partial z^2$)	kurtosis($\partial^2 T/\partial z^2$)
	$\max(\partial^2 T/\partial t^2)$	mean($\partial^2 T/\partial t^2$)	variance($\partial^2 T/\partial t^2$)	skewness($\partial^2 T/\partial t^2$)	kurtosis($\partial^2 T/\partial t^2$)
Other	cooling time	solidification time	v_1	v_2	v_3

Finally, batch sizes smaller than 4 yielded unstable training that failed to converge whereas batch sizes larger than 16 yielded models with tendencies to get stuck in local minima of the loss function and exhibit suboptimal prediction accuracy.

The remaining hyperparameter candidates were selected based on the success studies reported in previous literature. For example, LeakyReLU was chosen as an activation based on its improved performance over the previous gold standard ReLU activation as shown in previous literature (Maas et al., 2013). The swish function was chosen for similar reasons, as it too has demonstrated improved results over ReLU (Ramachandran et al., 2017). As Leaky ReLU and swish have not yet been comprehensively compared in the literature, we elected to investigate and compare them for the tensile strength prediction learning task in DED. A similar thought process was followed for the weight regularizers: $||\cdot||_1$ and $||\cdot||_2$ regularizers have both been shown to improve generalization performance of neural networks, i.e., by making training and testing error more consistent with one another, but there is not a definitive answer as to which one is best, so we elected to investigate both (Larsen and Hansen, 1994). Finally, Nadam and Adamax were previously shown to be the top two optimizers for learning tasks on large datasets, including fashion MNIST and natural images (Dogo et al., 2018). While the learning task and network architecture of this study is different, there is a lack of comparative research on optimizer behaviors for MLP models and so Nadam and Adamax were compared in this study.

To search over the hyperparameter ranges in a computationally tractable way, 5000 combinations are uniformly sampled from the hypercube defined by the nine hyperparameters of interest. In other words, discrete hyperparameters (#1, 2, 3, 4, 8, and 9) are independently sampled from discrete uniform distributions whereas the continuous hyperparameters (#5, 6, 7) are independently sampled from continuous uniform distributions. To avoid redundancy, none of the 5000 sampled combinations were identical. Each combination of hyperparameters is used to retrain the MLP and the hyperparameter combinations which yielded the highest average coefficient of determination (R^2) on the UTS and YS testing data are chosen for this study. Total MLP training time across all 5000 hyperparameter combinations was 166 h (average of approximately 2 min/model) on a single-threaded 2.6 GHz CPU. The MLP design and hyperparameter optimization process is illustrated in Fig. 6.

4. Results and discussion

The prediction outcomes of the ultimate tensile stress and yield stress for the 135 thin-walled thermal coupons of IN718 are analyzed using the method developed in this study.

4.1. Baseline prediction results

Baseline MLP prediction results for UTS and YS are shown in Fig. 7. The scatterplots in the top half of the figure are aggregated over 10 network reinitializations, each with stratified and randomly chosen testing data and randomly initialized weights. To quantify model performance and compare with results published in prior literature (Xie et al., 2021), which serves as benchmark for the comparison, four performance metrics are evaluated: root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and R^2 (Everitt and Skrondal, 2010). As evidenced by the similar bar heights in Fig. 7, the developed MLP has achieved comparable results to the benchmark, which was obtained by an 18-layer CNN and represents the best results in the current literature.

For a more rigorous comparison, two one-sided Welch's *t*-tests are performed on each performance metric to quantify performance differences (Everitt and Skrondal, 2010). The null hypothesis of the first test ($H^1{}_0$) is that the MLP performance is comparable to or *worse* than the CNN (i.e., higher RMSE, MAE, MAPE; lower R^2). The null hypothesis of the second test ($H^2{}_0$) is that the MLP performance is comparable to or *better* than the CNN. If H_0 is not rejected in either test, then the only possible explanation is that the MLP and CNN performances are statistically equivalent. However, if $H^1{}_0$ is rejected while $H^2{}_0$ is not, then the most likely explanation is that the MLP metric is statistically better than the CNN metric.

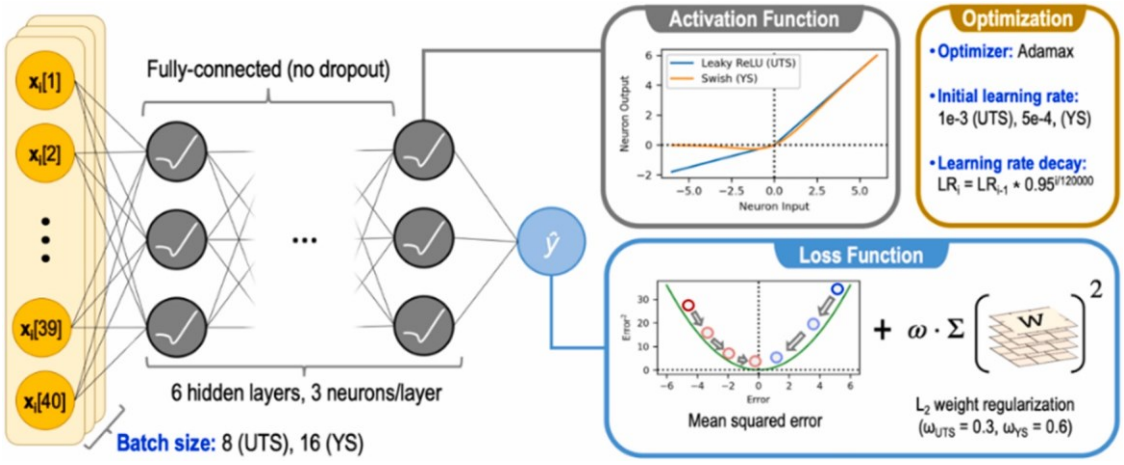

Fig. 6. Illustration of MLP design and hyperparameters optimization process.

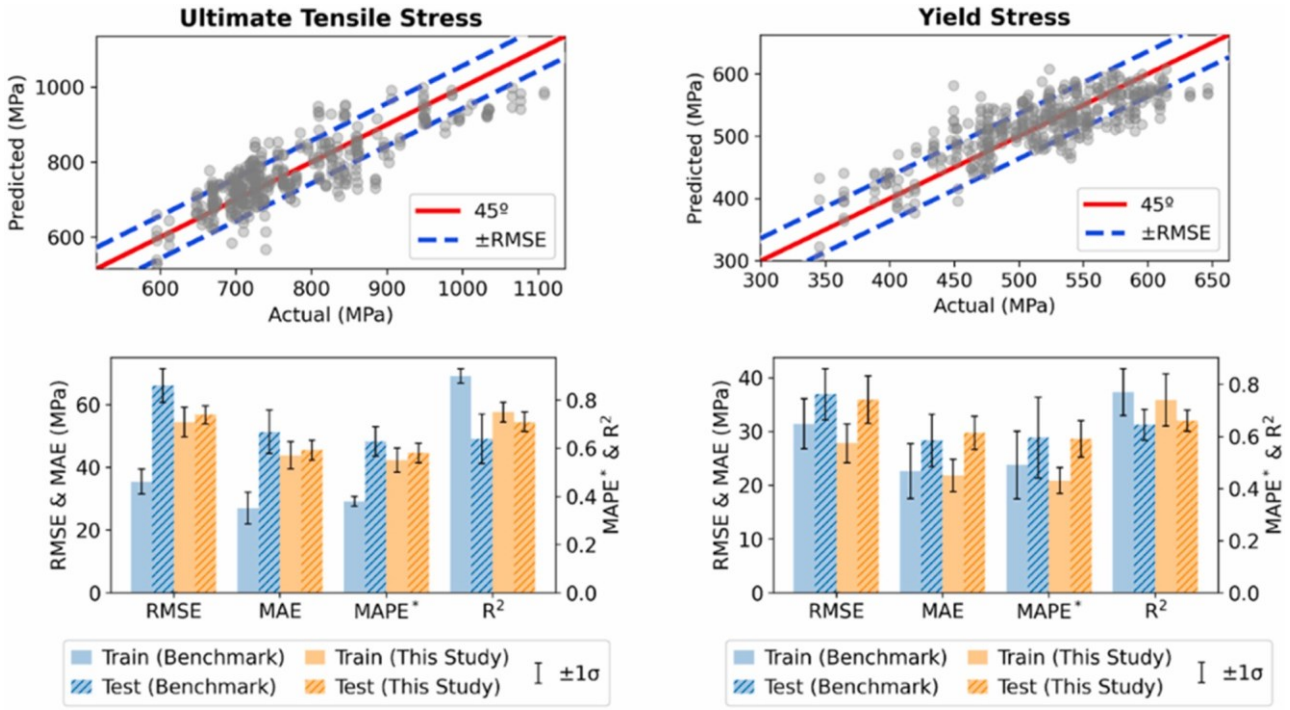

Fig. 7. Comparison of UTS and YS baseline models (MAPE* = MAPE × 10 for legibility); benchmark results are average of validation and testing results weighted by validation and testing proportions (0.44 and 0.56, respectively).

The Welch tests' significance level of $\alpha = 0.05$ means that there is a 5% chance of mistakenly rejecting H_0 when it is true. This rejection occurs when the p-value (p) is less than α, where p is the probability of observing the performance differences between the MLP and CNN assuming H_0 is true. A low p suggests that the observations are not well- explained by H_0 whereas a high p supports H_0. Welch's t-test is selected due to the different number of samples in each group (5 in benchmark vs. 10 in this study).

As seen in Table 2, all $p > 0.05$ for H^1_0 so there is not enough evidence to reject H^1_0 for any metric. Additionally, $p > 0.05$ for H^2_0 except for UTS RMSE, thereby confirming statistical equivalence between the MLP and CNN for all metrics except UTS RMSE, which is statistically better than the benchmark study with 99% confidence ($p_{UTS,2} = 0.01$).

In terms of computational efficiency, the MLP model requires 379 floating point operations (FLOPs) for a forward pass, as compared with 1.8×10^9 FLOPs for the ResNet-18 CNN (He et al., 2015) model used by the benchmark study. Therefore, for a fixed rate of FLOP/s, the MLP reduces computation time by a factor of more than 4.7×10^6. This is illustrated by the computational performance when predicting UTS and YS for all the 135 ROIs used in this study: the computational time needed for the prediction was approximately 1.2 µs for the MLP model when tested on a laptop with a single-threaded i7 CPU at 2.6 GHz. The time was about 5.8 s when using the CNN model. When scaling up the computation for analyzing larger ROI sample sizes, e.g., to characterize an entire DED built part, the difference in computational performance

becomes much more significant and affects the practicality of the implementation. These results clearly indicate that the MLP models developed in this study can achieve *comparable or better* predictive accuracy while being much *faster* and therefore more *scalable* to a DED production environment.

Table 2

Metric	$p_{UTS,1}$	Reject H^1_0?	$p_{UTS,2}$	Reject H^2_0?	$p_{YS,1}$	Reject H^1_0?	$p_{YS,2}$	Reject H^2_0?
RMSE	0.99	No	0.01	Yes	0.65	No	0.35	No
MAE	0.94	No	0.06	No	0.29	No	0.71	No
MAPE	0.92	No	0.08	No	0.53	No	0.47	No No
R_2	0.90	No	0.10	No	0.69	No	0.31	

Welch's one-sided *t*-test results for test data metrics at significance level α= 0.05.

4.2. Baseline model interpretation

For a physically consistent model, when a feature's value changes, the change should have a physically consistent effect on the tensile property. This is verified via the MLPs' SHAP values.

The SHAP values for the two baseline MLP models for the UTS and YS analysis are shown in Fig. 8. For each feature, the *x*-axis *location* and *color* of each dot represent the quantified *influence* of the feature and the normalized *value* of the feature itself, respectively. In addition, each dot represents a tensile coupon in the test dataset and there are 40 dots per row since each test dataset coupon yielded one SHAP value per feature. Features are shown in descending order of the SHAP CRV values, and Z- scores are used to normalize the differing scale of features.

As seen in Fig. 8, the top-three influential features for predicting UTS and YS are: 1) temperature variance within the truncated region (see Figs. 5), 2) the maximum temperature experienced by the DED part during the build process, and 3) the cooling time. Longer cooling times have been previously shown to increase the UTS and YS (Glerum et al., 2021). Additionally, solidification time is in the top five features for the UTS and YS models and is inversely proportional to both UTS and YS, as reflected in the coloration of the points along the solidification time rows in Fig. 8. These findings are again consistent with previous literature (Glerum et al., 2021). Furthermore, higher maximum temperature is seen to lead to low YS, which is consistent with reported literature where a positive correlation between maximum temperature and subsurface porosity was observed (Paulson et al., 2020). These results indicate that the feature influences learned by the MLP models from the data, without specialized architecture or training, are validated by physical relationships for UTS and YS prediction as reported in prior works.

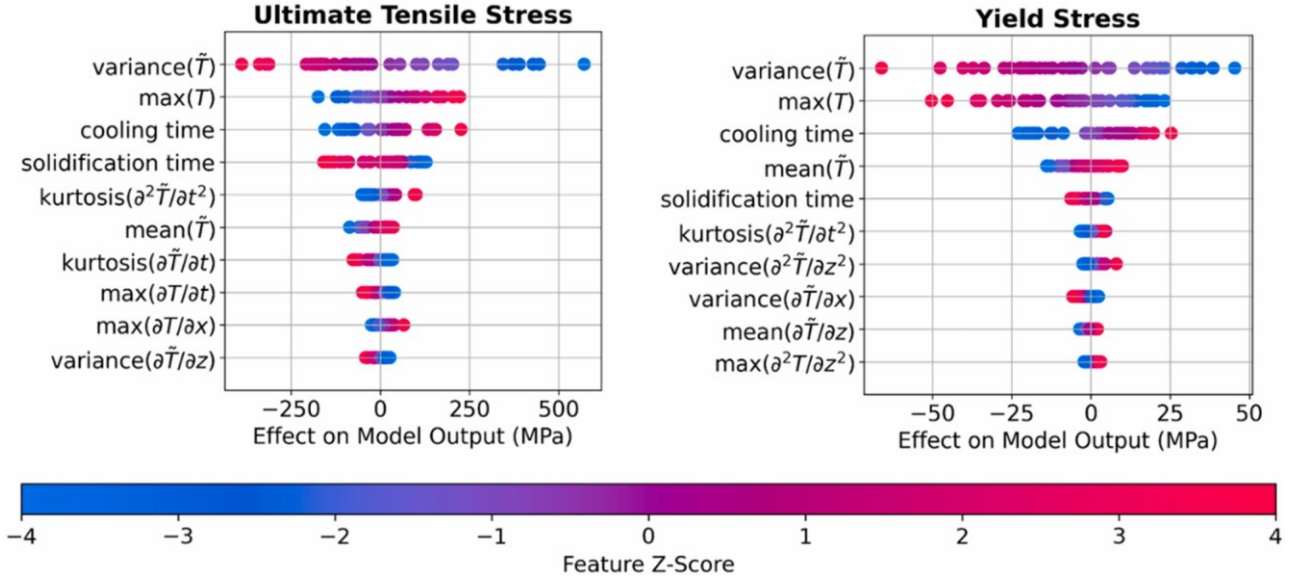

Fig. 8. Test dataset SHAP values for 10 most influential features of the MLP models with the highest R^2 values.

In addition to known physics, Fig. 8 also reveals new possible tensile strength predictors. For example, temperature variance is the most influential feature in both models and retains a consistent inverse proportionality to UTS and YS, which has not been reported previously. Similarly, UTS is observed to be directly proportional to the maximum temperature, which is an unreported phenomenon. While further study is necessary to reveal the underlying physical root causes of the feature rankings, many of the influential features identified by SHAP in Fig. 8 are supported by previous literature and thus lend credence to other unconfirmed predictors.

To better visualize how each SHAP-identified feature characterizes the tensile properties of the three deposition methods, SHAP values are color-coded according to their deposition type in Fig. 9. Two notable details emerge from these plots:

1) The mean and variance of the melt pool control specimens is consistently lower than the other two deposition methods. For the temperature variance feature in particular, the average normal, dwell, and melt pool control SHAP values in the UTS model are
− 153 ± 125 MPa, 253 ± 200 MPa, and − 102 ± 62.1 MPa, respectively, where the uncertainty reflects 1 standard deviation. For the YS model, these ranges are − 20.3 ± 18.9 MPa, 22.6 ± 17.5 MPa, and − 10.7 ± 6.4 MPa. For the UTS model, the temperature variance feature's SHAP value variance is 27813 for the normal and dwell data (on average) and 3856 for the melt pool control data. For the YS model, these variances are 332 and 41, respectively. In other words, the SHAP value variance of the normal and dwell models is, on average, 7.2 times larger for the UTS model and 8.1

times larger for the YS model. These variances, along with the lower average magnitude of the melt pool control SHAP values, suggests that the melt pool control predictions are reliant on more features than temperature variance, unlike the normal and dwell specimens which get most of the prediction "pushing power" from this one feature.

2) The normal and dwell SHAP values tend to exist on opposite sides of the plot for each feature, with melt pool control specimens occupying the middle. This suggests large statistical differences and easy distinguishability between the normal and dwell deposition types, which may be useful if automated deposition type recognition is required.

It should be noted that temperature variance, while having shown to provide a statistical predictor for tensile strength, is not physically validated as yet in this study. Specifically, different thermal histories can yield the same temperature variance, and the same temperature variance may lead to different tensile properties. Additionally, it is not yet established if temperature variance is a causal factor of tensile strength or rather an indicator of deposition type. As shown in Fig. 9, normal, melt pool control, and dwell specimen groups are discernable along the horizontal axis for both UTS and YS, indicating that the value of temperature variance discriminates between deposition types. Previous studies have shown that deposition type has a strong influence on tensile properties as it determines in-place heat treatment and precipitate formation characteristics, which are directly causal to tensile properties (Glerum et al., 2021). It is unknown if temperature variance encodes these characteristic differences or if temperature variance provides additional information beyond that of the deposition method. These findings indicate that further investigations involving physical models and specimen fabrication are needed to understand the role that temperature variance plays in affecting material properties and its potential utility as a tensile property predictor.

4.3. Feature removal and effects

The top row of Fig. 10 shows how SHAP CRV varies as a function of the number of inputs to the UTS and YS models. Each MLP was retrained using the minimum number of features necessary to achieve the CRV thresholds as specified by the legends in Fig. 10. These thresholds were representatively chosen to expedite the search for an optimal feature set, as discussed in Section 2.3. The steep curves in the top row of Fig. 10 show that most of the SHAP value variance can be explained with only a few features in both models. This is an indicator that significant feature pruning is possible to improve the computational efficiency and model interpretability. Feature pruning is illustrated in the bottom row of Fig. 10, which shows that model performance on the test dataset is largely unaffected by feature removal, with 100% and 75% SHAP CRV models differing in performance by metrics no more than 10%.

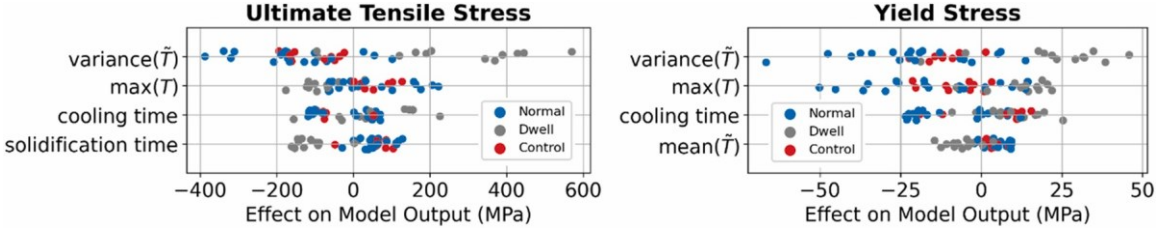

Fig. 9. SHAP values color-coded per deposition type for top 4 most influential features of MLP model; vertical jitter has been added for visual clarity.

To determine which model of Fig. 10 has the optimal performance-interpretability tradeoff, (2) is solved via the additive utility function defined in (7). As such, the metrics of interest are defined as **m** = [RMSE, MAE, MAPE, R^2, number of input features] with the weights being η = [.125,.125,.125,.125,.50] to evenly balance performance metrics with parsimony via a 50/50 weighting scheme. The selected error metrics (RMSE, MAE, MAPE, R^2) were considered such that the outcome of this study can be compared with the published literature (Xie et al., 2021) to demonstrate the improvement of the developed method, whereas the number of input features was selected for the utility function such that model parsimony could be balanced against model performance, which is one of the main motivators of this study.

Eq. (8) is then used to find **m'** for each metric based on the results shown in Fig. 10. The utility values, U, for the UTS and YS models are shown in Table 3. They indicate that 75% is the optimal SHAP CRV threshold for both models since the utility is maximized.

The 75% SHAP CRV threshold corresponds to 3-feature and 2-feature models for UTS and YS, respectively. Based on the feature rankings in Fig. 8, the optimal UTS model uses variance(T), max(T), and cooling time as its inputs and the optimal YS model uses only variance(T) and max(T), leading to two *parsimonious* predictive models with good model performance. These features are identified since they are at the top of each plot in Fig. 8, which is sorted in descending order of RV. Since the optimal UTS and YS feature sets were made using the CRV heuristic, it is known that each pruned model's features are selected in descending order of RV and thus the features at the top of Fig. 8 were chosen before any others.

Using the 3-feature model, UTS prediction metrics are shown to be worsened by 6.0% on average whereas YS metrics *improved* by an average of 5.6% using the 2-feature model. These performance alterations are accompanied by 92.5% and 95% reductions in input feature set size, respectively. The UTS result is expected and agrees with machine learning literature, which suggests a tradeoff between interpretability and performance (Dziugaite et al., 2020). On the other hand, the YS result is an indicator of a better fitted YS model. The discrepancy between the training and testing YS metrics was 1/4th as large as the full YS model in Fig. 7. The pruned YS model therefore experienced less overfitting than the full-feature model and thus is more robust when predicting never-before-seen YS data, such as the test dataset. Such a reduction in overfitting is attributable to the reduced input size of the pruned model, which resulted in 114 fewer free parameters (− 30%) and significantly less overparameterization when predicting YS.

The performance of the 3- and 2-feature MLPs for UTS and YS is compared against 40 different regression models including k-nearest neighbors, boosted regressors, bagged regressors, tree- and forest-based models, support vector machines, Gaussian process, generalized linear models, and step-wise methods such as LASSO, ridge, ElasticNet, LARS, and orthogonal matching pursuit regression. The full list is shown in (Shankar, 2022). Each model is fit on the features used to train the maximally parsimonious MLP. After the initial fitting process, the top 5 models, as measured by the average R^2 across UTS and YS, are fine tuned to maximize their predictive accuracy, and then compared to the MLP results. As shown in Fig. 11, none of the fine-tuned models are shown to outperform the MLP in terms of R^2. This confirms the strength of the developed prediction model as compared to other modelling approaches.

5. Conclusions

The presented study aims to fill an existing research gap in interpretable process-property predictive modeling of DED. An efficient feature extraction methodology and corresponding machine learning models have been developed to exploring the tensile strength- determining physics embedded in the DED thermal history. Experimental data analysis has demonstrated that, compared with prior published results as the benchmark, the developed MLP models have comparable or better prediction accuracy for DED built IN718 UTS and

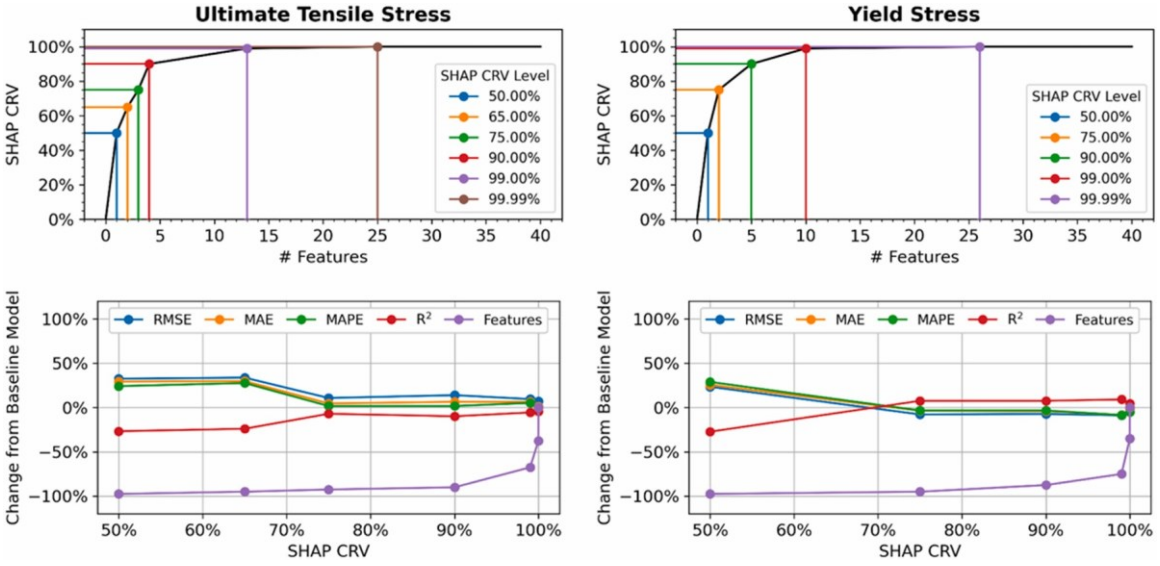

Fig. 10. Cumulative relative variance plots (top) and effects of feature removal on test dataset results (bottom) for UTS and YS models.

Table 3

Standardized performance metrics (h), weights (η), utility (U), and rankings for MLP models at varying SHAP CRV levels, evaluated on the test dataset.

	Ultimate Tensile Stress							Yield Stress						
						U	Rank	$m'[1]$					U	Rank
SHAP CRV	$m'[1]$	$m'[2]$	$m'[3]$	$m'[4]$	$m'[5]$				$m'[2]$	$m'[3]$	$m'[4]$			$m'[5]$
50.00%	0.04	0.01	0.13	0.00	1.00	0.52	5	0.00	0.00	0.00	0.00	1.00	0.50	5
65.00%	0.00	0.00	0.00	0.11	0.97	0.50	T6	N/A	N/A	N/A	N/A	N/A	N/A	N/A
75.00%	**0.68**	**0.84**	**0.94**	**0.74**	**0.95**	**0.87**	**1**	**0.97**	**0.86**	**0.86**	**0.96**	**0.97**	**0.96**	**1**
90.00%	0.59	0.79	0.94	0.63	0.92	0.83	2	0.95	0.87	0.86	0.96	0.90	0.92	2
99.00%	0.72	0.79	0.81	0.79	0.69	0.73	3	1.00	1.00	1.00	1.00	0.77	0.90	3
99.99%	0.80	0.96	1.00	0.84	0.38	0.64	4	0.87	0.86	0.91	0.87	0.36	0.64	4
100.00%	1.00	1.00	1.00	1.00	0.00	0.50	T6	0.72	0.75	0.77	0.75	0.00	0.39	6
η	0.125	0.125	0.125	0.125	0.5	-	-	0.125	0.125	0.125	0.125	0.5	-	-

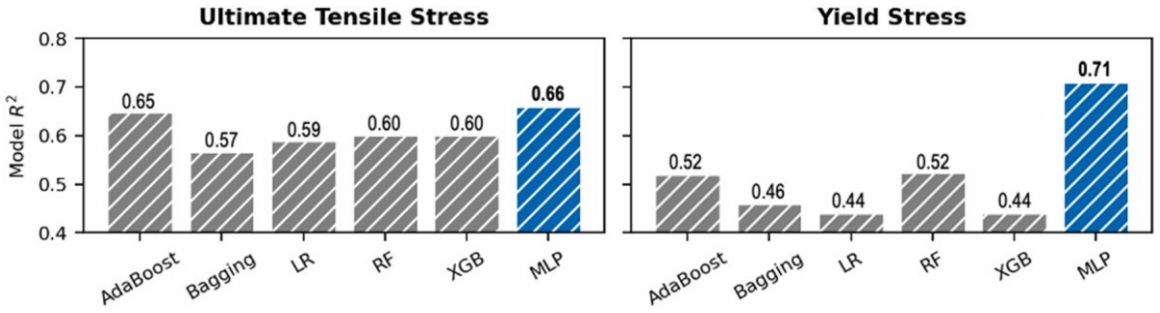

Fig. 11. Comparison of this study's parsimonious MLP and comparable statistical models showing that the MLP yields the highest R^2 value; LR: linear regression, RF: random forest, XGB: extreme gradient boosting; the AdaBoost (Freund and Schapire, 1995), Bagging (Aslam et al., 2008), RF (Breiman, 2001), and XGB (Chen and Guestrin, 2016) ensemble methods each used decision trees as their ensemble members.

YS with significantly reduced model complexity (by a 4.7 ×10^6 reduction factor). This complexity reduction was made possible by a novel feature pruning method. SHAP values of the model input features were investigated to confirm the models' consistency with physical principles underlying the DED process as reported in published literature. Note that this work uses the discovery of the importance of cooling time on tensile strength obtained from the neural network in Xie et al. as one feature, and the analysis of this work confirms its relevant importance. The success of this work indicates that similar knowledge from the traditional heat treatment field can be and should be used in feature definition to reduce model complexity in physics-informed machine learning.

Additionally, previously unreported predictors such as temperature variance are observed in the presented study, which can enable new DED optimization strategies. For instance, temperature variance has been shown to be a possible predictor of tensile strength, pending further experimental verification. Furthermore, the robustness of the developed methods to changes in machine settings and material will also need to be analyzed.

More investigation into the SHAP CRV-based feature selection method is also needed. For brevity, the presented study did not explore possible inefficiency induced by reducing the number of possible feature subsets from $2^{|F|}$ to $|F|$ to $N \ll |F|$. Furthermore, why feature removal made the YS MLP perform better than the 40-feature full model, as seen in Fig. 10, needs further investigation. These tasks will be necessary to fully understand the method's behavior. Finally, a comprehensive, microstructural interpretation of every variable remaining in the reduced models is needed. This will help gain insight into the DED physics in the context of established material science and synthesize additional relationships between thermal history and tensile properties that may not have been revealed yet. Overall, the presented study constitutes a significant step towards physics-informed and parsimonious

machine learning modeling of DED and other metal powder-based additive manufacturing processes.

CRediT authorship contribution statement

Clayton Cooper: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Jianjing Zhang:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Robert X. Gao:** Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Jennifer Bennett:** Data curation, Experimentation, Resources, Writing – review & editing. **Jian Cao:** Funding acquisition, Experimentation, Project administration, Resources, Supervision, Writing – review & editing. **Joshua Huang:** Software, Validation, Writing – review & editing.

Declaration of Competing 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.

Data Availability

The data that has been used is confidential.

Acknowledgements

CC acknowledges support by the National Science Foundation Graduate Research Fellowship under Grant No. 1937968. JB and JC acknowledge support by the Army Research Laboratory under grant W911NF-18-2-0275. JC, RG, and JZ acknowledge support by the National Science Foundation under ERC (HAMMER) grant EEC-2133630.

References

Aslam, J.A., Popa, R.A., Rivest, R.L., 2008. On Estimating the Size and Confidence of a Statistical Audit (Working Paper No. 54), Voting Technology Project. Caltech/MIT.

Bai, S., Kolter, J.Z., Koltun, V., 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv:1803.01271 [cs].

Bennett, J., Kafka, O., Liao, H., Wolff, S., Yu, C., Cheng, P., Hyatt, G., Ehmann, K., Cao, J., 2018. Cooling rate effect on tensile strength of laser deposited Inconel 718. Procedia Manuf. 26, 912–919. https://doi.org/10.1016/j.promfg.2018.07.118.

Bennett, J., Glerum, J., Cao, J., 2021. Relating additively manufactured part tensile properties to thermal metrics. CIRP Ann. https://doi.org/10.1016/j.cirp.2021.04.053.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Callister, W.D., Rethwisch, D.G., 2012. Fundamentals of materials science and engineering: an integrated approach, 4th ed. ed. Wiley, Hoboken, N.J.

Chen, J., Yuan, S., Lv, D., Xiang, Y., 2021. A novel self-learning feature selection approach based on feature attributions. Expert Syst. Appl. 183, 115219 https://doi.org/10.1016/j.eswa.2021.115219.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Presented at the KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, San Francisco California USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Clarke, K.D., 2014. Austenite Formation and Microstructural Control in Low-Alloy Steels. Comprehensive Materials Processing. Elsevier, pp. 345–361. https://doi.org/ 10.1016/B978-0-08-096532-1.01211-5.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Math. Control Signal Syst. 2, 303–314. https://doi.org/10.1007/BF02551274.

Dass, A., Moridi, A., 2019. State of the art in directed energy deposition: from additive manufacturing to materials design. Coatings 9, 418. https://doi.org/10.3390/coatings9070418.

Dogo, E.M., Afolabi, O.J., Nwulu, N.I., Twala, B., Aigbavboa, C.O., 2018. A Comparative Analysis of Gradient Descent-Based Optimization Algorithms on Convolutional Neural Networks, in: 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). Presented at the 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), IEEE, Belgaum, India, pp. 92–99. https://doi.org/10.1109/ CTEMS.2018.8769211.

Dozat, T., 2016. Incorporating Nesterov Momentum into Adam, in: Proceedings of the 33rd International Conference on Machine Learning. Presented at the International Conference on Machine Learning, San Juan, PR, p. 4.

Dziugaite, G.K., Ben-David, S., Roy, D.M., 2020. Enforcing Interpretability and its Statistical Impacts: Trade-offs between Accuracy and Interpretability. arXiv: 2010.13764 [cs, stat].

Everitt, B., Skrondal, A., 2010. The Cambridge dictionary of statistics.

Farshidianfar, M.H., Khajepour, A., Gerlich, A.P., 2016. Effect of real-time cooling rate on microstructure in Laser Additive Manufacturing. J. Mater. Process. Technol. 231, 468–478. https://doi.org/10.1016/j.jmatprotec.2016.01.017.

Freund, Y., Schapire, R.E., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In: Vitanyi, P. (Ed.), Computational Learning Theory, ´ Lecture Notes in Computer Science. Springer Berlin Heidelberg, Berlin, pp. 23–37. https://doi.org/10.1007/3-540-59119-2_166.

Gibson, I., Rosen, D., Stucker, B., 2015. Directed energy deposition processes. Additive Manufacturing Technologies. Springer New York, New York, NY, pp. 245–268. https://doi.org/10.1007/978-1-4939-2113-3_10.

Glerum, J., Bennett, J., Ehmann, K., Cao, J., 2021. Mechanical properties of hybrid additively manufactured Inconel 718 parts created via thermal control after secondary treatment processes. J. Mater. Process. Technol. 291. https://doi.org/10.1016/j.jmatprotec.2021.117047.

Gockel, J., Beuth, J., Taminger, K., 2014. Integrated control of solidification microstructure and melt pool dimensions in electron beam wire feed additive manufacturing of Ti-6Al-4V. Addit. Manuf. 1–4, 119–126. https://doi.org/10.1016/ j.addma.2014.09.004.

Guo, S., Agarwal, M., Cooper, C., Tian, Q., Gao, R.X., Guo, W., Guo, Y.B., 2022. Machine learning for metal additive manufacturing: towards a physics-informed data-driven paradigm. J. Manuf. Syst. 62, 145–163. https://doi.org/10.1016/j. jmsy.2021.11.003.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs].

Hejripour, F., Binesh, F., Hebel, M., Aidun, D.K., 2019. Thermal modeling and characterization of wire arc additive manufactured duplex stainless steel. J. Mater. Process. Technol. 272, 58–71. https://doi.org/10.1016/j.jmatprotec.2019.05.003.

Hocking, R.R., 1976. A biometrics invited paper. the analysis and selection of variables in linear regression. Biometrics 32, 1. https://doi.org/10.2307/2529336.

Kingma, D.P., Ba, J., 2017. Adam: A Method for Stochastic Optimization. arXiv: 1412.6980 [cs].

Kou, S., 2003 (ed.). Welding Metallurgy, Second ed. Wiley-Interscience, Hoboken, N.J.

Larsen, J., Hansen, L.K., 1994. Generalization performance of regularized neural network models, in: Proceedings of IEEE Workshop on Neural Networks for Signal Processing. Presented at the IEEE Workshop on Neural Networks for Signal Processing, IEEE, Ermioni, Greece, pp. 42–51. https://doi.org/10.1109/NNSP.1994.366065.

Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems 30. Presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, California.

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models, in: Proceedings of the 30th International Conference on Machine Learning. Presented at the International Conference on Machine Learning, Atlanta, GA.

Malakooti, B., 2014. Operations and Production Systems with Multiple Objectives. John Wiley & Sons Inc, Hoboken, New Jersey.

Marcilio, W.E., Eler, D.M., 2020. From explanations to feature selection: assessing SHAP values as feature selection mechanism, in: 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). Presented at the 2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE, Recife/Porto de Galinhas, Brazil, pp. 340–347. https://doi.org/10.1109/SIBGRAPI51738.2020. 00053.

Mokdad, F., Bouchaffra, D., Zerrouki, N., Touazi, A., 2015. Determination of an optimal feature selection method based on maximum Shapley value, in: 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA). Presented at the 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA), IEEE, Marrakech, Morocco, pp. 116–121. https://doi. org/10.1109/ISDA.2015.7489211.

Moran, M.J. (Ed.), 2003. Introduction to Thermal Systems Engineering: Thermodynamics, Fluid Mechanics, and Heat Transfer. Wiley, New York.

Ness, K.L., Paul, A., Sun, L., Zhang, Z., 2021. Towards a generic physics-based machine learning model for geometry invariant thermal history prediction in additive manufacturing. J. Mater. Process. Technol., 117472 https://doi.org/10.1016/j. jmatprotec.2021.117472.

Paulson, N., Gould, B., Wolff, S., Stan, M., Greco, A., 2020. Correlations between thermal history and keyhole porosity in laser powder bed fusion. Addit. Manuf. 34. https://doi.org/10.1016/j.addma.2020.101213.

Qin, J., Hu, F., Liu, Y., Witherell, P., Wang, C.C.L., Rosen, D.W., Simpson, T.W., Lu, Y., Tang, Q., 2022. Research and application of machine learning for additive manufacturing. Addit. Manuf. 52, 102691 https://doi.org/10.1016/j.addma.2022.102691.

Ramachandran, P., Zoph, B., Le, Q.V., 2017. Searching for Activation Functions. arXiv: 1710.05941 [cs].

Roscher, R., Bohn, B., Duarte, M.F., Garcke, J., 2020. Explainable machine learning for scientific insights and discoveries. IEEE Access 8, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199.

Salman, S., Liu, X., 2019. Overfitting Mechanism and Avoidance in Deep Neural Networks. arXiv:1901.06566 [cs, stat].

Sanchez, S., Rengasamy, D., Hyde, C.J., Figueredo, G.P., Rothwell, B., 2021. Machine learning to determine the main factors affecting creep rates in laser powder bed fusion. J. Intell. Manuf. 32, 2353–2373. https://doi.org/10.1007/s10845-021- 01785-0.

Seifi, M., Salem, A., Beuth, J., Harrysson, O., Lewandowski, J.J., 2016. Overview of materials qualification needs for metal additive manufacturing. JOM 68, 747–764. https://doi.org/10.1007/s11837-015-1810-0.

Shankar, P., 2022. LazyPredict.
Shapley, L., 1951. Notes on the N-Person Game – II: The Value of an N-Person Game. RAND Corporation. https://doi.org/10.7249/RM0670.
Snow, Z., Reutzel, E.W., Petrich, J., 2022. Correlating in-situ sensor data to defect locations and part quality for additively manufactured parts using machine learning. J. Mater. Process. Technol. 302, 117476 https://doi.org/10.1016/j.jmatprotec.2021.117476.
Sun, H., Mukherjee, S., Daly, M., Krishnan, A., Karigerasi, M.H., Singh, C.V., 2016. New insights into the structure-nonlinear mechanical property relations for graphene allotropes. Carbon 110, 443–457. https://doi.org/10.1016/j.carbon.2016.09.018.
Vafadar, A., Guzzomi, F., Rassau, A., Hayward, K., 2021. Advances in metal additive manufacturing: a review of common processes, industrial applications, and current challenges. Appl. Sci. 11, 1213. https://doi.org/10.3390/app11031213.
Xie, X., Bennett, J., Saha, S., Lu, Y., Cao, J., Liu, W.K., Gan, Z., 2021. Mechanistic data- driven prediction of as-built mechanical properties in metal additive manufacturing. npj Comput. Mater. 7, 86. https://doi.org/10.1038/s41524-021-00555-z.
Young, H.P., 1985. Monotonic solutions of cooperative games. Int. J. Game Theory 14, 65–72. https://doi.org/10.1007/BF01769885.
Zhang, Z., Liu, Z., Wu, D., 2020. Prediction of melt pool temperature in directed energy deposition using machine learning. Addit. Manuf. https://doi.org/10.1016/j. addma.2020.101692.
Zoga ła-Siudem, B., Jaroszewicz, S., 2021. Fast stepwise regression based on multidimensional indexes. Inf. Sci. 549, 288–309. https://doi.org/10.1016/j.ins.2020.11.031.