



Additive Manufacturing Melt Pool Prediction and Classification via Multifidelity Gaussian Process Surrogates

Robert Saunders^{1,2} · Anna Rawlings¹ · Andrew Birnbaum¹ · Athanasios Iliopoulos¹ · John Michopoulos¹ · Dimitris Lagoudas² · Alaa Elwany²

Received: 13 June 2022 / Accepted: 26 August 2022

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2022

Abstract

It is well known that the process parameters chosen in metal additive manufacturing (AM) are directly related to the melt pool dimensions, which can be related to microstructure characteristics, properties, and printability. Thus, the determination of melt pool dimensions and resulting printability, given a set of process parameters, is crucial to understanding the performance of AM parts. Unfortunately, experiments relating process parameters to the melt pool have a high execution and data collection cost. On the other hand, simulations do not suffer this cost penalty but can never be as good as the ground truth experiments. A number of capabilities exist to predict melt pool geometry from process parameters that range from highly accurate codes with long simulation times to less accurate analytical solutions that are nearly instantaneous. This work leverages multifidelity Gaussian process (GP) surrogates to examine how the fusion of information from different fidelities, including experiments, influences the resulting predictive model. Both a multifidelity GP regression and a novel multifidelity GP classification are examined. The multifidelity models are compared to standard GPs and a withheld set of test data. The results show good predictive accuracy for the melt pool width and depth, but suggest multifidelity models may not significantly improve classification of printability.

Keywords Metal additive manufacturing · Melt pool prediction · Printability · Surrogate modeling

Introduction

Additive manufacturing (AM) remains an area of critical technological and strategic importance to the US government (USG) and its allies [1]. While AM has been actively researched as a science and engineering discipline for the past decade, its adoption as a mainstream manufacturing technology, particularly for metals, in practice has been slow, in part due to a lack of understanding of the performance of as-built AM parts and the challenges with qualification/certification that stem from this lack of understanding [2]. The Office of the President of the USA recently set out to address this lack of understanding in AM with AM Forward, a partnership between the USG and the manufacturing

industry to conduct research to improve the performance of AM techniques [3]. While this effort primarily focuses on lowering the cost and increasing the quality and adoption of AM parts in industry, there are additional USG-wide efforts to utilize AM technologies to create a more secure and resilient supply chain [4–6]. In addition to these efforts, the US Department of Defense has identified AM as a crucial technology needing further development for US national security with the ability to create lighter, stronger, and more resilient components to better prepare and protect Warfighters [7].

In order to improve the performance of metallic AM parts, the type of defects that arise during manufacturing and the causation mechanisms for the defects must be better understood [8]. The process to manufacture a metal AM part typically involves a metal powder being spread in a powder bed or deposited through a nozzle, and then an energy source (such as a laser or electron beam) is used to melt the powder in specific locations within a layer guided by a CAD file [9]. The molten material then cools and solidifies to create the desired geometry of a layer. This process is repeated until a full part is complete. The repeated thermal

✉ Robert Saunders
robert.saunders@nrl.navy.mil

¹ U.S. Naval Research Laboratory, Washington, DC 20375, USA

² Department of Industrial & Systems Engineering, Texas A&M University, College Station, TX 77840, USA

cycling of the powder and underlying solidified material in conjunction with the high thermal gradients and material phase transformation can lead to parts containing numerous defects that, ultimately, degrade the parts performance [8]. A key aspect to understanding the defects that are formed is understanding how the AM process parameters influence the resulting melt pool, or the liquid interface between the powder particles and the energy source [10, 11]. Too little energy being deposited will result in a melt pool not being formed, while a high energy will result in unsteady melt pool dynamics that can have unintended consequences [2]. In both cases, porosity defects will form. In the no melting case, a lack of fusion defect can form and lead to a lower effective material density part with irregular-shaped voids. In the high energy case, defects, such as keyholing and balling/beading, can lead to entrapped gas and again result in a part with a lower effective material density. In order to avoid these defects, the so-called *printability* needs to be assessed. Printability in this work is defined as the process parameter map that can tell whether or not a porosity defect (*i.e.*, lack of fusion, keyholing, balling/beading) will occur for a given set of process parameters [12–17]. One popular way to assess and create this printability map is by *in-situ* process monitoring [18]. By monitoring the process and collecting the data, machine learning (ML) models can be trained to predict thermal history [19–22], melt pool size and shape [15], anomalies [23], and printability [16, 17]. However, this work involves the continued collection of experimental data, which can be costly and time-consuming, in order to train the ML models.

A less-costly alternative is to utilize one or more of the plethora of available modeling techniques to simulate every level of the AM process [24, 25]. Many of these models focus solely on predicting the thermal history and corresponding melt pool geometry as functions of input process parameters [26]. These process models range from analytical [27], and semi-analytical methods [28–32], to finite element/difference models [14, 33–39], to multiphysics models at the scale of the metal powder [40–46]. They generally fall into a spectrum from fast and approximate (*e.g.*, analytical solutions) to slow and accurate (*e.g.*, multiphysics finite element or powder-scale models). In all cases, there will be some level of inaccuracy and uncertainty introduced by the modeling assumptions with higher fidelity models generally having fewer simplifying assumptions [47]. With lower fidelity models, uncertainty quantification (UQ) techniques can be used to interrogate the model relatively easily due to the low data acquisition cost. With higher fidelity models, this is not practical and many researchers implement machine learning (ML) or surrogate modeling techniques, such as Gaussian process (GP) models, to conduct UQ tasks, such as model calibration, uncertainty propagation, or sensitivity analysis [14, 48–56]. GPs, in general, are a popular

tool to emulate AM process models as they limit the number of simulations required to effectively conduct UQ and can provide reasonably accurate approximations to the model they are emulating at a nearly negligible computational cost once trained.

A popular variation of the standard GP is the multifidelity GP developed by Kennedy and O'Hagan [57], commonly referred to as the Kennedy-O'Hagan or KOH model. In this framework, multiple sources of information at varying degrees of fidelity can be incorporated into a single model. This limits the number of expensive simulations and/or experiments needed to build a model by leveraging information from cheap low fidelity solutions to the same problem. One example of this is using a few experiments to adjust a computer simulation [58, 59]. Recently, Mahmoudi et al [50] applied this same methodology to a finite element (FE) AM process model to perform calibration of the model parameters and were able to achieve a high predictive accuracy.

In this work, a new multifidelity GP (MFGP) approach based on experiments and multiple fidelities of simulations is presented. Two MFGPs are shown, one that performs regression to determine melt pool dimensions and another that performs classification to determine printability. The MFGPs are trained using a combination of experiments and simulations for a model material of 316L stainless steel, and the data generation is scaled based on the fidelity (*i.e.*, a few experiments, but many analytical solution data points). In doing so the MFGP is made to be as efficient as possible when compared to a standard single fidelity GP. The remainder of this work will be structured as follows: “[AM Process Modeling](#)” section overviews the AM process models utilized to generate data; “[Eagar-Tsai Analytical Solution](#)” section provides a brief overview of the MFGP regression along with a novel MFGP classification approach; “[NRL-Enriched Analytical Solution Method](#)” section explains the data generation process from the models as well as how the regression and classification models are trained; “[Finite Element Modeling](#)” section presents the validation of the MFGP against a test set of experiments, demonstrates how different combinations of fidelities can yield the best overall model, and analyzes how each process model contributes to the overall MFGP; and finally “[Experimental Data](#)” section summarizes the work and provides prospects for future directions.

AM Process Modeling

This work utilizes four different fidelity information sources to generate the data needed to train the MFGPs. These are, namely, the analytical solution of Eagar and Tsai [27], the Naval Research Laboratory (NRL) enriched analytical

solution method [28, 29], a conduction-based FE model implemented in COMSOL Multiphysics® [60], and laser powder bed fusion (L-PBF) experiments. The Eagar–Tsai and NRL enriched analytical solution method are both briefly detailed below and the interested reader is referred to the referenced works for full details. The COMSOL FE model and experiments are detailed in full as they are both previously unpublished. The computational models are setup to simulate the L-PBF process with 316L stainless steel as used in the experiments. Additionally, each model takes as input the 316L material properties and three process parameters, namely the laser power, velocity, and spot size¹. The boundary value problem for the simulations is such that each considers a semi-infinite domain where the laser source is far from any boundaries. The top surface is a free surface, and mirror symmetry is implemented where applicable for half of the laser source and melt pool width. These conditions mimic those of the experiments where each single track is far from the base plate edges and, while multiple tracks are on a single plate, each is far from the others and a small dwell time is used to ensure temperature effects from previous tracks are not influencing the current track. Note that none of the simulations are specifically calibrated to the experiments and only use material parameters obtained from literature for a general 316L material. Temperature fields from each simulation are used to extract melt pool dimensions once a steady state thermal solution is achieved. Due to the high data acquisition and analysis cost of the higher fidelity information and limitations of the lowest fidelity model, only single-track simulations are considered to maintain consistency between the fidelities.

Eagar–Tsai Analytical Solution

The Eagar–Tsai (ET) analytical solution is the lowest fidelity source used in this work and calculates the temperature field due to a traveling Gaussian distributed heat source on a semi-infinite plate and uses the calculated temperature field to compute melt pool dimensions. The solution is based upon fundamentals of heat transfer, albeit with a number of simplifying assumptions, such as non-temperature-dependent material properties and only heat conduction physics with no radiative or convection losses. The ET solution was originally developed to describe the impact of welding process parameters on the geometry and temperature distribution of weld melt pools [61]. However, it can be readily applied to develop a first-order approximation for temperatures in the L-BPF process since L-PBF can be thought of

as a repetitive micro-welding process [17, 62, 63]. While the ET solution is not highly accurate for predicting all aspects of the L-PBF process, it is extremely fast and takes very little computational resources thus providing a good starting point for further analysis. A single simulation can run in seconds on a single i7 (tenth generation) CPU in a laptop.

NRL-Enriched Analytical Solution Method

The NRL-enriched analytical solution method (NEASM) was developed by modifying the ET solution to incorporate several enrichments to better capture the features seen in the AM process. These are: 1) using the linear heat equation to approximate the nonlinear heat equation solution to account for the temperature-dependent properties, via a fixed point iteration, 2) using the method of images to account for finite domains, 3) inclusion of mass accretion by accounting for mass conservation, and 4) including phase transformation. Enrichments 1 and 4 allow for temperature-dependent materials to be included in the NEASM, while enrichments 2 and 3 account for realistic part geometries and effects of multiple layers, respectively. In general, the NEASM is capable of predicting the thermal history of full AM part builds including rastering within a layer and the effects of multiple layers. However, for this work, only single-track data are needed; thus, it can be assumed that boundary effects are negligible (*i.e.*, the laser is sufficiently far from all boundaries and enrichment 2 is not necessary) [64]. Additionally, under certain circumstances, the NEASM can produce results with accuracy comparable to finite element modeling [29]. Each simulation in this work using the NEASM takes under 1 minute on a single i7 (tenth generation) CPU in a laptop.

Finite Element Modeling

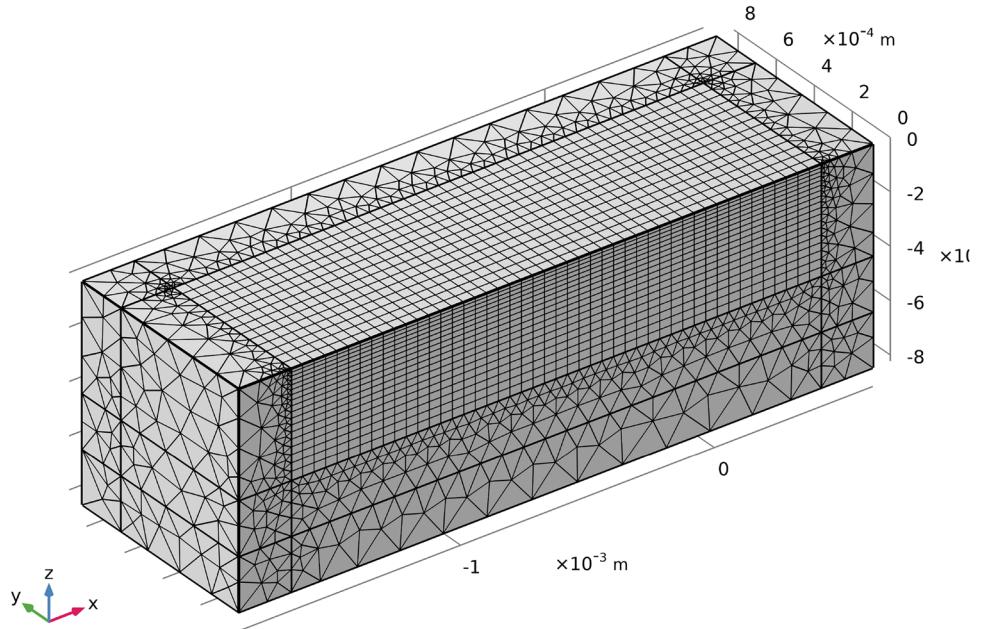
The finite element analysis (FEA) parametric simulation was implemented in COMSOL Multiphysics® 5.6 [60], in conjunction with the LiveLink™ capability enabling direct communication between COMSOL and MATLAB® [65]. The MATLAB scripting capability offers the opportunity to automate the generation of the datasets required to construct the surrogate models. The developed simulation capability is based on an advective Eulerian approach that significantly reduces the computational cost as compared to a traditional Lagrangian approach. The solved equation expresses the local balance of energy and reduces to the steady-state form of the advective heat conduction equation:

$$\rho(T)C_p(T)\mathbf{u} \cdot \nabla T - \nabla \cdot \mathbf{q} = 0, \quad (1)$$

with T the temperature, $\mathbf{q} = -k(T)\nabla T$, the heat flux vector defined by the Fourier constitutive equation, $\rho(T)$ the temperature-dependent density, $C_p(T)$ the temperature-dependent

¹ More specifically, spot size is used to denote the distance equivalent to four standard deviations of a Gaussian beam profile, commonly referred to as D4σ

Fig. 1 One realization of the parameterized, discretized domain used in the COMSOL heat transfer simulations



heat capacity, $k(T)$ the temperature-dependent thermal conductivity, $\mathbf{u} = \{-v\mathbf{i}, 0\mathbf{j}, 0\mathbf{k}\}^T$ the velocity vector, and v the velocity magnitude of the deposited beam power source in the x -direction (applied to the material domain, while the heat source is maintained fixed in space).

One realization of the discretized domain is shown in Fig. 1. The xz plane at $y = 0$ is considered a symmetry plane by enforcing $-\mathbf{n} \cdot \mathbf{q} = 0$, with \mathbf{n} the normal vector of the surface. The domain is discretized with tetrahedral elements, utilizing the feature of COMSOL Multiphysics[®] for enabling semi-infinite domains with the intent to simulate an arbitrarily large geometries. A convective heat flux boundary condition was applied using $-\mathbf{n} \cdot \mathbf{q} = h(T_\infty - T)$ to simulate the presence of material at the outside of these domains. T_∞ was taken as the room temperature. The value of h was set by simulating a complete domain with adequate material, such that T_∞ and T were equal, then comparing that simulation to a reduced domain that implemented the convective boundary conditions. After iterating, a value of h was determined that simulates the convection corresponding to the case with 316L material being the medium on the other side of the wall. The approximate value identified was $h = 500 \text{ W}/(\text{m}^2\text{K})$. It should be noted that the results of the simulation are very insensitive to the actual value of h since the infinite element domain already addresses simulating boundary conditions at very large distances. The deposited beam power was applied at the top boundary of the hexahedral elements in the form of a heat flux boundary condition given by:

$$-\mathbf{n} \cdot \mathbf{q} = \alpha P_0 f(\mathbf{O}, \mathbf{e}) \frac{|\mathbf{e} \cdot \mathbf{n}|}{\|\mathbf{e}\|}, \quad (2)$$

with $\alpha = 0.45$ the laser coupling coefficient, P_0 the laser power and $\mathbf{e} = \{0\mathbf{i}, 0\mathbf{j}, -1\mathbf{k}\}^T$ the beam orientation vector. The function f defines the deposited beam shape and was assumed to be of a Gaussian form according to:

$$f(\mathbf{O}, \mathbf{e}) = \frac{1}{2\pi\sigma^2} e^{-\frac{s^2}{2\sigma^2}}, \quad (3)$$

with

$$s = \frac{\|\mathbf{e} \times (\mathbf{x} - \mathbf{O})\|}{\|\mathbf{e}\|}, \quad (4)$$

where \mathbf{x} is the vector of coordinates of each boundary point and \mathbf{O} is the center of the beam application, considered here to be at point $\{0, 0, 0\}^T$. The Gaussian distribution variance was assumed to be equal to $d/4$ with d the apparent beam diameter, i.e., $D4\sigma$ as defined previously.

One challenge with developing the automated dataset generation relates to the very wide extent of the parametric values, which entails different geometry sizes for different values of the parameters because of the different spread in temperature distribution. Since it is impossible to know *a priori* the extent of the melt pool, an iterative algorithm was developed to detect when the geometry needed to be adjusted to account for the size of the melt pool. The algorithm performs the simulation using nominal values, and if it identifies that the width or length of the melt pool is very close to the boundaries, it extends the domain and executes another simulation. The process is repeated until the

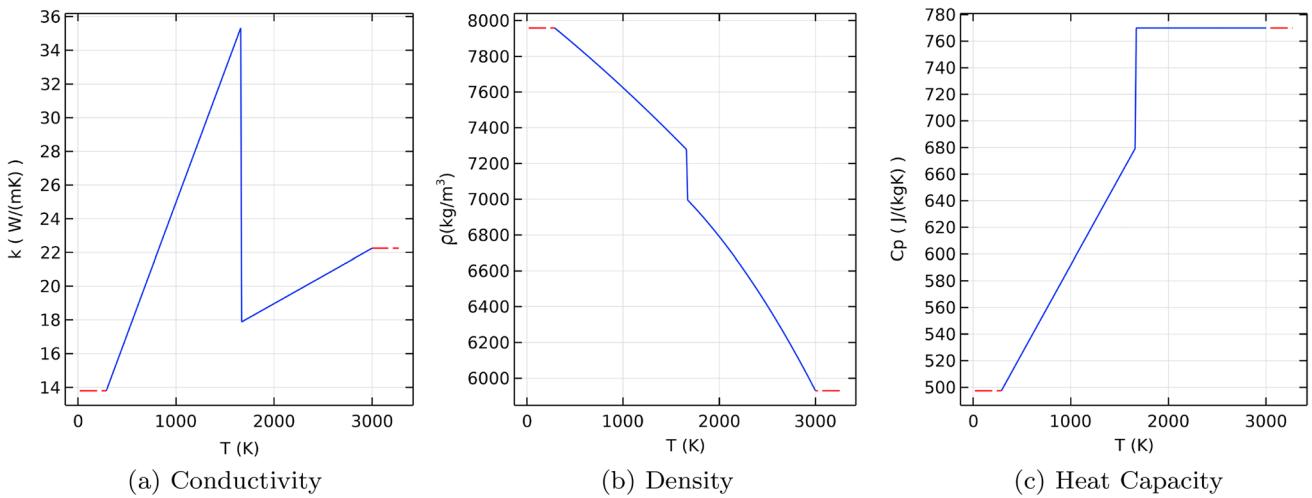


Fig. 2 Temperature-dependent material properties assumed for the heat transfer problem

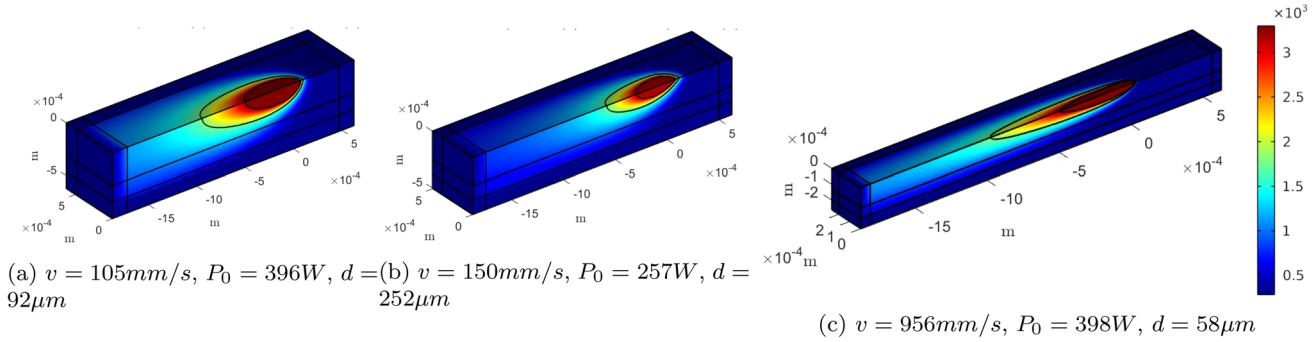


Fig. 3 Example temperature distributions for three different parameter combinations

computational domain is at least twice as big as the extent of the melt pool in all three directions.

The simulations were performed using the temperature-dependent material properties shown in Fig. 2. The extent of the melt pool in the three directions representing width, length, and depth were calculated by finding the extrema of the temperature iso-surfaces for $T = T_{melt} = 1660 \text{ K}$. Three simulations showing temperature distributions and domain size changes with parameters are shown in Fig. 3. The computational time for each of the simulations on a 4-core i7 (tenth generation) laptop was at the level of 20–25 s.

Experimental Data

Experiment (EXP) data used for training and testing the MFGP scheme are single-track, so-called “bead-on-plate”/autogenous, experiments conducted on a GE/Concept Laser M2 selective laser melting (SLM) system on a 316L stainless steel plate [66]. The nominal processing conditions for the system are 900 mm/s laser velocity, 370 W laser power,

and 160 μm laser (2σ) spot size. The minimum and maximum ranges of the system are 10–7000 mm/s, 75–400 W, and 50–350 μm . These ranges are used to bound the data generation in the experiments as well as the other models with the exception that a maximum laser velocity of 2000 mm/s is set. This velocity was chosen as a practical maximum given the maximum possible power of the system *i.e.*, without an increase in power, such high velocities will seldom result in melted material.

Melt pool width and depth were extracted from the single-track experiments using optical microscopy. The samples were mounted and subjected to standard stainless steel grinding and polishing procedures (*i.e.*, 320 grip paper, 9 μm diamond polishing, 3 μm diamond polishing, then 40 nm OP-S polishing). 10% oxalic acid electroetching at 5V for 15 s was used to reveal the melt pool. Once mounted and processed, optical microscopy was performed. Melt pool width and depth were measured from the optical images. A representative output image from this process is shown in Fig. 4a. In addition to melt pool width and depth, the melt

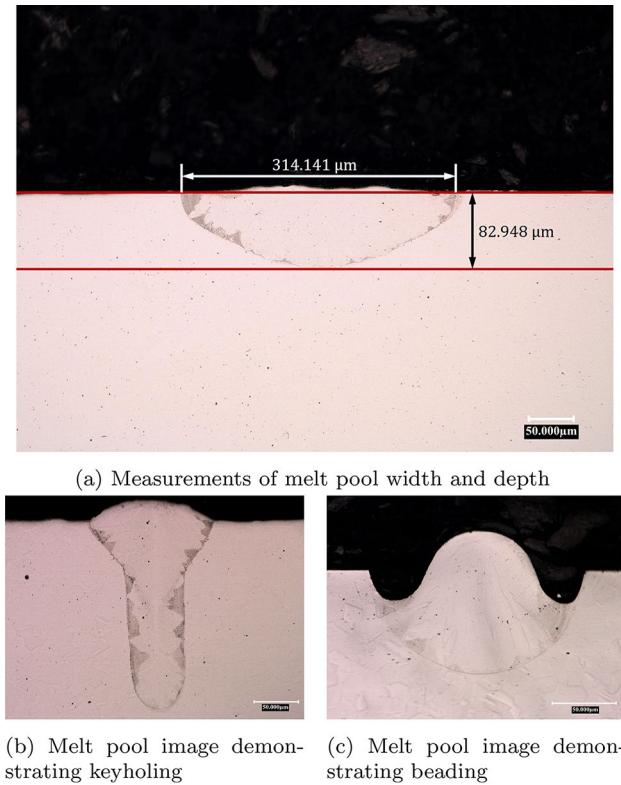


Fig. 4 Representative single-track experiment SEM images

pools were visually evaluated to determine if the run was in keyholing or conduction mode (Fig. 4b), whether beading (Fig. 4c), or if there was lack of fusion. Note that melt pool depth was measured from the surface in all cases for consistency.

Multifidelity Gaussian Processes

The KOH model [57] has been demonstrated to be very effective at multifidelity information fusion over the past two decades. However, it has two primary limitations. First, there is a linear correlation assumed between the higher fidelity data and the next lowest fidelity data. This linear correlation assumption is generally adequate but, in many computer models, this linear correlation only holds for specific ranges of the model. For instance, the ET model of this work always assumes conduction is the dominant heat transfer mechanism. When this is true, the ET model will likely be linearly correlated to higher fidelity models. However, at process parameters where conduction is not the primary mechanism, other models accounting for multiple heat transfer mechanisms could give a significantly different prediction that is not linearly correlated to the ET solution. The second limitation of the KOH model is that the computational complexity to train

the model scales as $\mathcal{O}((\sum_{t=1}^s n_t)^3)$ where s is the number of fidelities and n_t is the number of data points at each fidelity. This computational complexity can quickly become intractable considering that the lowest fidelity models can easily generate hundreds or thousands of points. However, both of the stated limitations have been overcome in recent years with Le Gratiet et al. [67, 68], addressing the latter and Perdikaris et al. [69], addressing the former. In the following derivations, standard GP regression and classification theory is briefly mentioned but for complete details, the interested reader is directed to the book of Rasmussen and Williams [70].

First, assume that there is a dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\} = (\mathbf{x}, y)$ for $i = 1, \dots, n$ with input vectors \mathbf{x}_i and responses y_i for n points. These data are assumed to be generated by some unknown latent function $f(\cdot)$ which follows an n -dimensional multivariate Gaussian distribution such that,

$$p(f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)) \sim \mathcal{N}_n(\boldsymbol{\mu}, \mathbf{k}), \quad (5)$$

where $\boldsymbol{\mu}$ is the mean vector defined by the mean function $\mu(\mathbf{x}_i) = \boldsymbol{\mu}_i = \mathbb{E}[f(\mathbf{x}_i)]$ and \mathbf{k} is the covariance defined by the covariance function $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{k}_{ij} = \text{cov}[f(\mathbf{x}_i), f(\mathbf{x}_j)]$. Now, the unknown latent function can be assigned a Gaussian process prior denoted as $f(\cdot) \sim \mathcal{GP}(\mu(\cdot), k(\cdot, \cdot))$. As is common, throughout this work, it will be assumed that this and all other GP priors are zero mean *i.e.*, $\mu(\cdot) = 0$.

In the context of multifidelity modeling, there are now multiple datasets from each fidelity such that $\mathcal{D}_t = (\mathbf{x}_t, y_t)$ for $t = 1, \dots, s$ with the s -level being the highest fidelity and the first level being the lowest. In the KOH auto-regressive model, the s level are correlated as

$$f_t(\mathbf{x}) = \rho f_{t-1}(\mathbf{x}) + \delta_t(\mathbf{x}), \quad (6)$$

where ρ is a scaling factor that linearly correlates the t and $t - 1$ fidelities and the bias of the lower fidelity is captured by $\delta_t(\mathbf{x}) \sim \mathcal{GP}(0, k_t(\cdot, \cdot))$. As mentioned above, the construction of the model in this way has a cost of $\mathcal{O}((\sum_{t=1}^s n_t)^3)$ due to the required inversion of the covariance matrix to compute the posterior. Le Gratiet et al. derived a more numerically efficient scheme by replacing the GP prior $f_{t-1}(\mathbf{x})$ with the corresponding posterior $f_{t-1}^*(\mathbf{x})$ while maintaining $\mathcal{D}_t \subseteq \mathcal{D}_{t-1}$. Since nested data (*i.e.*, data from fidelity t is a subset of the points in fidelity $t - 1$) have been assumed, the posterior predictive density $f_{t-1}^*(\mathbf{x})$ is deterministic at \mathbf{x} , which essentially decouples the MFGP problem into s standard GP problems. It was shown that this scheme results in the exact same posterior as the KOH model, while being more efficient (computational cost of $\mathcal{O}(\sum_{t=1}^s (n_t^3))$) and yielding predictive models for each fidelity rather than only the highest fidelity as is the case in the KOH model.

To account for nonlinear correlations, Perdikaris et al. modified equation 6 as

$$f_t(\mathbf{x}) = z_{t-1}(f_{t-1}(\mathbf{x})) + \delta_t(\mathbf{x}), \quad (7)$$

where $z_{t-1}(\cdot)$ is an unknown function that maps the lower fidelity to the higher fidelity. This unknown function is assigned a GP prior such that $z_{t-1}(\mathbf{x}) \sim \mathcal{GP}(0, k_t(\cdot, \cdot))$. By assigning a GP prior to $z_{t-1}(\cdot)$, the posterior of $f_t(\cdot)$ is no longer Gaussian and is considered a “deep GP.” These deep GPs are generally very computationally complex, but can be made more tractable by following the scheme of Le Gratiet et al. where the prior of fidelity $t - 1$ is replaced by the posterior as

$$f_t(\mathbf{x}) = z_{t-1}(f_{t-1}^*(\mathbf{x})) + \delta_t(\mathbf{x}) = g_t(\mathbf{x}, f_{t-1}^*(\mathbf{x})), \quad (8)$$

where $g_t(\mathbf{x}) \sim \mathcal{GP}(0, k_t((\mathbf{x}, f_{t-1}^*(\mathbf{x})), (\mathbf{x}', f_{t-1}^*(\mathbf{x}'))))$. This formulation is made possible by the independence of $z_{t-1}(f_{t-1}^*(\mathbf{x}))$ and $\delta_t(\mathbf{x})$ as well as the fact that the sum of two GPs results in another GP. Perdikaris et al. note that this formulation retains the equivalent Markov property of the KOH model and the scheme of Le Gratiet et al. . Furthermore, with certain covariance kernel choices, the formulation of Le Gratiet et al. can be obtained. In this work, the covariance function of g_t takes the decomposed form of

$$k_{t_g} = k_{t_p}(\mathbf{x}, \mathbf{x}') \cdot k_{t_f}(f_{t-1}^*(\mathbf{x}), f_{t-1}^*(\mathbf{x}')) + k_{t_\delta}(\mathbf{x}, \mathbf{x}'), \quad (9)$$

where each k_t is a valid covariance function. The covariance functions will take the common form of a stationary, squared exponential covariance as

$$k_t(\mathbf{x}, \mathbf{x}') = \eta_t \exp\left(-\frac{1}{2} \sum_{k=1}^p \theta_{k,t} (x_k - x'_k)^2\right). \quad (10)$$

When $k_{t_\delta}(\mathbf{x}, \mathbf{x}')$ takes the form as specified, it will result in a nonlinear correlation, but in portions of this work, a linear correlation will be used for comparison and that will result in $k_{t_\delta}(\mathbf{x}, \mathbf{x}')$ taking the form of a bias or constant kernel. The parameters $\{\eta_t, \theta_{1,t}, \dots, \theta_{p,t}\}$ for $t = 1, \dots, s$ make up the set of so-called hyper-parameters, which allow “tuning” of the correlation between data points and fidelities, and are learned from the data at each fidelity along with the posterior of the previous fidelity.

Note that up this point, no differentiation has been made between regression and classification. In both settings, the unknown latent function has the form as given and the posterior distribution can be found as

$$p(f_t^*(\mathbf{x}^*)) = \int p(f_t(\mathbf{x}^*, f_{t-1}^*(\mathbf{x}^*)) | y_t, \mathbf{x}_t, \mathbf{x}^*) p(f_{t-1}^*(\mathbf{x}^*)) d\mathbf{x}^*. \quad (11)$$

In the standard GP regression problem and the nonlinear auto-regressive GP (NARGP) regression with $t = 2$, the likelihood is Gaussian and along with the assumed Gaussian priors, the posterior predictive distribution is Gaussian, and

the integral can be computed analytically. However, NARGP regression with $t > 2$ and in all GP classification, the posterior is not Gaussian and alternative methods must be used to compute the integral to obtain the posterior predictive distribution.

Regression

Even though the posterior for the NARGP with more than two levels is not Gaussian, the process to compute it is still quite simple. First, each level of the model is trained individually. At the first level, this corresponds to a standard GP regression with $\mathcal{D}_1 = (\mathbf{x}_1, y_1)$. Subsequent levels require modifying the training dataset such that $\mathcal{D}_t = ((\mathbf{x}_t, f_{t-1}^*(\mathbf{x}_t)), y_t)$. Acquiring the posterior $f_{t-1}^*(\mathbf{x}_t)$ does not require approximating the integral of the previous fidelity, because the posterior of fidelity $t - 1$ is deterministic at all points \mathbf{x}_t due to the assumed nested structure of the data. This training process is repeated for all fidelities until the s -levels are trained. Once trained, the posterior predictive density at a new point \mathbf{x}^* can be found at all fidelities $t > 2$ by using Monte Carlo (MC) integration of equation 11. For fidelity 1 and 2, the posterior can be analytically obtained since it is Gaussian. Note that using MC integration requires the propagation of all MC samples from the previous fidelity to the current fidelity, while MC sampling again at the current fidelity for each of the MC samples generated at the previous fidelity. In doing so, the number of MC samples required to approximate the posterior scales exponentially and at the highest fidelity requires $(n_{MC})^{s-2}$ total MC samples.

Classification

Unlike GP regression, the posterior for GP classification always requires approximation methods to evaluate the posterior. In GP classification, the latent function is considered a nuisance function where its value is never observed and, once a value is determined, it is transformed through a non-linear warping to make predictions of observations. For the case of binary classification, which this work will focus on, this results in $\pi(\mathbf{x}) \triangleq p(y = 1 | \mathbf{x}) = \Phi(f(\mathbf{x}))$, where $\pi(\cdot)$ is a deterministic function and, in this work, $\Phi(\cdot)$ is the standard Normal cumulative distribution function *i.e.*, a probit function is used for the warping. The probit function has the purpose of warping the latent function from $[-\infty, \infty]$ to $[0, 1]$, which is then used as a prediction of the class probability. Since this work focuses only on binary classification, the data likelihood will take the form of a Bernoulli distribution. [71]. For the case of a single fidelity classifier, methods to approximate the posterior have been well-studied and the approximation can be done via Laplace approximation or expectation propagation (EP) [70], or MC methods such as

Markov chain Monte Carlo (MCMC). However, multifidelity GP classification has not been as well-studied. Recently, Sahli Costabal *et al.* [72] developed and demonstrated a multifidelity GP classifier that was based on the KOH model and used MCMC to approximate the posterior. Likewise, Klyuchnikov and Burnaev [73] presented a Laplace approximation method for Gaussian process classification with two fidelities. However, this again was based on the KOH model using a linear correlation structure.

In this work, a novel approach to MFGP classification is taken. In general, all equations derived up to this point still hold. However, each dataset $\mathcal{D}_t = (\mathbf{x}_t, y_t)$ now has labels y_t take the form of binary classes 0 or 1, rather than continuous variables. This allows the predictive posterior for the data to be written as

$$\pi_t^*(\mathbf{x}^*) \triangleq p(y^* = 1 | y_t, \mathbf{x}_t, \mathbf{x}^*) = \int \Phi(f_t^*(\mathbf{x}^*)) p(f_t^*(\mathbf{x}^*)) d\mathbf{x}^*. \quad (12)$$

One can note that this posterior is the same as the posterior given in Rasmussen and Williams [70], except generalized to multiple fidelities and including a dependence (via equation 11) on the previous fidelity posterior prediction. By maintaining the assumptions given above, the problem of MFGP classification is reduced to fitting s single fidelity classifiers as developed by Le Gratiet *et al.* [67, 68], and allows the nonlinear correlation structure developed by Perdikaris *et al.* [69], to be maintained *i.e.*, this approach is nearly identical to the NARGP regression above. The primary differences between the NARGP regression and classification are in training. First, to train each fidelity classifier, one must implement an approximation method as discussed above. In this work, that approximation is made using EP. Using an approximate inference method like EP is useful here as it allows the computational efficiency to be maintained. If implementing an MC type approach, sampling s fidelities can quickly become cumbersome and costly. From here, the approach to training is exactly as it is in the case of regression where each fidelity is trained recursively and the posterior of the latent function is used as an input to the next fidelity. It is important to emphasize here that, while the latent function in classification is unobserved, it does provide information about the underlying relationship between the data. If posterior predictive distribution for the data is used here as the input the next fidelity, the warping function will tend to map points that could be very far apart so that they appear close together and related. Once the models are trained, MC integration is performed exactly as is done in the regression case except posteriors for both the latent function and the data are needed. The latent function posterior samples are propagated to the next fidelity, while the posterior of the data is the prediction of the GPC for the current fidelity.

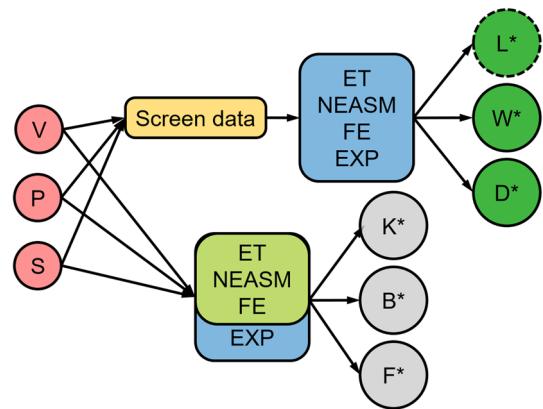


Fig. 5 MFGP regression and classification framework. Process parameter inputs are shown at the left, these are fed into the NARGP regression (possibly after some data screening, discussed later) and classification, which are used to create predictions of melt pool dimensions and printability. For regression, melt pool length is only trained for NARGP models without EXP. For classification, empirical classes are used for all simulation data

Data Generation and Model Training

In order to generate data to train the MFGP regression and classification, each of the four information sources detailed above must be sampled. The laser power (P), velocity (V), and spot size (S) are varied, while the melt pool length (L), width (W), depth (D), conduction/keyhole mode class (K), beading class (B), and lack of fusion class (F) are output. The ranges for the input parameters are based on the experimental setup and are detailed in “Experimental Data” section. Due to experimental limitation, melt pool length is not measured and is only available for simulations *i.e.*, the three lowest fidelities. Conversely, the simulations do not contain sufficient physics to produce conduction/keyhole mode classifications. In lieu of direct class outputs for the simulations, empirical measures based on melt pool dimensions will be used to generate the needed output [14]. Since these empirical measures require melt pool dimensions, obtained through observation or regression, classifiers built using the empirical measures will be referred to as regression-based classifiers. In this work, keyholing will correspond to a melt pool width to depth ratio of $W/D < 1.5$. Additionally, lack of fusion and beading classification will be predicted as $L/W > 2.3$ and $D/t < 1$, respectively, where t is the powder layer thickness, which will be $25 \mu\text{m}$. The set of these 3 classes will determine the so-called “printability” and give a predicted process parameter space where “good” (*i.e.*, defect free) printing should occur. The framework for MFGP regression and classification is shown in Fig. 5.

In the case of a single fidelity data source, data are typically generated via a space-filling design of experiments (DoE) method such as Latin hypercube sampling (LHS).

However, simply sampling each of the thermal models used here from the different fidelities would result in multiple space filling designs that would almost certainly not be nested, thus violating the assumptions of the theory above. With multifidelity modeling becoming more common in the past decade, numerous space filling nested DoE (nDoE) methods have been developed [67, 68, 74–77]. In this work, the relatively simple scheme of Le Gratiet *et al.* [67, 68], is chosen. In this nDoE method, a space filling LHS design of n_s points is first created for the highest fidelity model. Subsequently, another LHS design is constructed for fidelity $s - 1$ with n_{s-1} points. From the n_{s-1} points, the points that are closest in space to the higher fidelity n_s points are removed and replaced by the points n_s . In this work, a simple Euclidean distance measure is used but any desired distance measure could equivalently be implemented. The process as described is recursively applied until LHS designs are generated and modified for all s -fidelities. For this work, the ET model is sampled 500 times, the NEASM sampled 200 times, FE sampled 100 times, and experiments (henceforth referred to as EXP) sampled 50 times. For training, all data are used except 15 experiments, which were randomly selected from the 50 available. The 15 experiments not used for training are used for validation/testing of the trained GPs.

The NARGP regression and classification as well as the nDoE are all implemented in Python 3.6.12. The NARGP regression and classification are based on Emukit (<https://emukit.github.io/>) [78], although the implementation has been modified to perform classification and cross-validation (although not used for this work) in the regression setting.

Results and Discussion

In this work, the NARGP will be trained on all combinations of fidelities with 2 or more fidelities included. In addition, each of the combinations will be trained once with a nonlinear correlation between fidelities and a second time assuming a linear correlation between fidelities. This will result in 22 trained NARGP models (Table 1), which will be compared to standard GPs trained on a single fidelity. Each of the trained models will first be compared to the experiment test set and discussed, followed by a comparison of which fidelities included in the NARGP result in the best models for the least computational cost will be discussed.

Experiment Validation

Regression

The purpose of regression is to determine steady-state melt pool dimensions as a function of the AM process parameters. Since a classification is being trained to screen for

Table 1 Multifidelity model combinations with corresponding labels of 1-22. Each design uses a binary labeling where a 1 indicates inclusion in the model and 0 exclusion. The nonlinear column specifies if a nonlinear correlation was used (1) or was not used (0)

	Nonlinear	EXP	FE	NEASM	ET
1	1	1	1	1	1
2	0	1	1	1	1
3	1	0	1	1	1
4	0	0	1	1	1
5	1	1	0	1	1
6	0	1	0	1	1
7	1	0	0	1	1
8	0	0	0	1	1
9	1	1	1	0	1
10	0	1	1	0	1
11	1	0	1	0	1
12	0	0	1	0	1
13	1	1	0	0	1
14	0	1	0	0	1
15	1	1	1	1	0
16	0	1	1	1	0
17	1	0	1	1	0
18	0	0	1	1	0
19	1	1	0	1	0
20	0	1	0	1	0
21	1	1	1	0	0
22	0	1	1	0	0

defect inducing parameters and the purpose of the regression model is to predict good melt pool dimensions that can be propagated to subsequent models, the regression model can be trained in one of two ways. First, utilizing all data including data which produces melt pools that exhibit keyholing, beading, and/or lack of fusion. Or second, screen the training data beforehand and only training the regression model on the most relevant good print data. In the former case, more data will be available to train the model, but the data may be unnecessary since some regions will never be queried for predictions and won't necessarily reduce the uncertainty in relevant good regions. For the sake of comparison, models are trained using both data sets in this work. In both instances, a GP will also be trained on the highest fidelity model used in the NARGP. Melt pool length is not included here as melt pool length measurements were not able to be collected in the experiments. Mean absolute percent error (MAPE) and coefficient of determination (R^2) are used to evaluate the error of the trained model on the test data. The way in which MAPE is formulated can sometimes be problematic as it can result in erroneous, infinite, or undefined values [79]. As such, both R^2 and MAPE are used initially with R^2 being used as a verification of the MAPE results. The results of training can be seen in Fig. 6a and c and

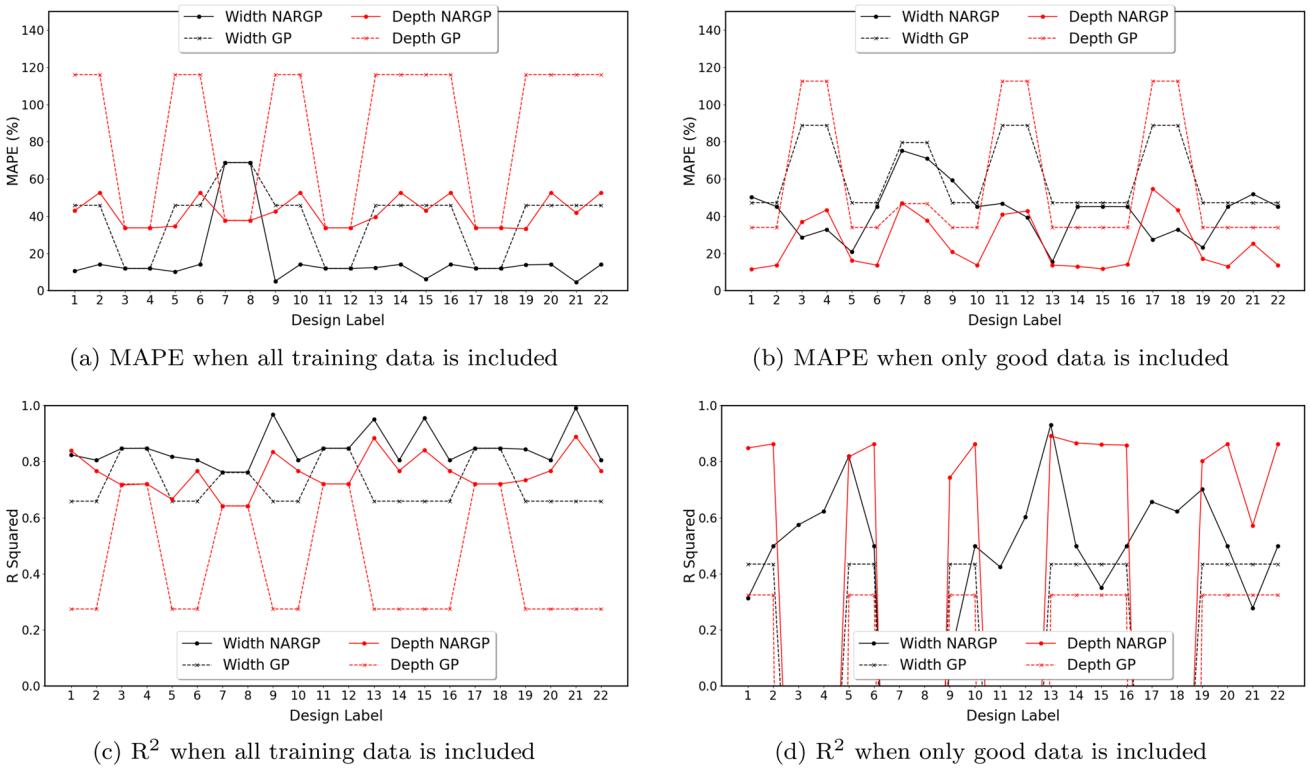


Fig. 6 MAPE and R^2 evaluated on test set melt pool width (W) and depth (D) data for the NARGPs of Table 1 and GPs

Fig. 6b and d for the models trained on all data and models trained on only good print data, respectively. Note, that in all cases, only test data that does not produce a defect will be used to evaluate the model. Since the regression model will only be used to predict good prints, these filtered test data will be the most relevant to evaluate true model accuracy. The filtered data set has seven test points remaining of the original 15.

From Fig. 6a and c, it can be seen that the GP can achieve a minimum error of 11.9% for the melt pool width and 33.7% for the depth when it is trained using the FE data. In general, the NARGP achieves a substantial improvement in accuracy, using both the MAPE and R^2 metrics, compared to the GP and in the worst cases, achieves a similar level of accuracy. In the best case scenario the NARGP can predict the melt pool width with 4.5% error and the depth with 33.7% error. For the melt pool width, the best model combinations were the EXP and FE with a nonlinear correlation structure, while for the depth, the best was any model trained using FE without experiments. The R^2 metric predicts a maximum melt pool width R^2 of 0.991 for the NARGP and 0.847 for the GP with a maximum melt pool depth R^2 of 0.889 for the NARGP and 0.72 for the GP. This again demonstrates that the NARGP produces a better model than the GP. The best melt pool width model using R^2 was again a combination

of EXP and FE with a nonlinear correlation structure. For melt pool depth, the R^2 and MAPE metrics differ on which combination of models produce the best prediction indicating that there may be a few bad predictions of the melt pool depth that are skewing the MAPE metric. The highest R^2 value attained for melt pool depth was 0.889 for the combination of EXP and FE with a nonlinear correlation structure.

For the GP only fitted to good data (Fig. 6b and 6 d), a similar level of accuracy as the GP trained on all data is achieved. The GP achieves a minimum error of 47.2% for the melt pool width and 33.9% for the depth when it is trained using the EXP data. As before and as expected, the NARGP improves on or maintains the accuracy and R^2 values achieved by the GP alone with the best case predicting the melt pool width with 15.6% error ($R^2 = 0.931$) using EXP and ET and a nonlinear correlation and the melt pool depth with 11.5% error ($R^2 = 0.848$) using all fidelities and a nonlinear correlation. Similarly, a melt pool depth error of 11.6% ($R^2 = 0.861$) can be achieved with the same fidelities as the previous except excluding the ET data. Interestingly, the best R^2 metric for melt pool depth of 0.891 was achieved using EXP and ET and a nonlinear correlation, which as before demonstrates that the best melt pool width and depth models determined by R^2 were produced using the same combination of models.

Regardless of the removal of data before training, the NARGP produces better results than a GP trained on a single fidelity. Furthermore, the NARGP demonstrates its value over traditional multifidelity GP modeling approaches as the nonlinear correlation between fidelities produces the models with the highest accuracy. As one might expect, the NARGP including experiments produces the models with the lowest error and highest R^2 values. However, this is not the case for the GP where the model using FE data alone can produce a more accurate model (in percent error) than the one using experiments. This further highlights the need for multifidelity modeling to effectively increase the amount of information available for model training because, intuitively, experiments should always be used when available as they represent the ground truth. However, due to their high acquisition cost, only a limited number of experiments may be available and when taken alone may not produce the most accurate model as has been demonstrated.

Looking at the difference in the best cases when using all data versus only good data, it can be noted that the single best melt pool width prediction is obtained using all data, while the best melt pool depth prediction is obtained when using the screened good data considering MAPE alone as a metric. When R^2 is used an additional metric, the same conclusion can be drawn that the screened data generally produces a better model but the difference in the best result is not as noticeable *e.g.*, the best R^2 for all data is 0.889 and for good data only, it is 0.891. This phenomenon can be attributed to how defects show up in the melt pool width and depth predictions. Melt pool width alone does not give a strong indication of a defect, unlike melt pool depth which can be a good indicator of keyholing and lack of fusion. Therefore, the melt pool width can be predicted more accurately when more data is available during training as is typically the case with GP-based models. For melt pool depth, the inclusion of data with defects and the fusion of information from models without the physics necessary to produce those defects tends to result in a model that is less accurate than a model trained on a consistent set of data *i.e.*, one where simulations and experiments are both operating under the same physics. A similar trend on the difficulty of predicting melt pool depth can be seen in Tapia *et al.* [51].

In the best cases, an error rate of around 5–10% could be achieved in prediction (slightly under for 5% for melt pool width and slightly over 10% for melt pool depth). Likewise, a very strong correlation can be achieved for both the melt pool width ($R^2 = 0.991$) and depth ($R^2 = 0.891$). Depending on the application this could be an acceptable level of accuracy given the trade-off with computational time. Some recent works [50, 53, 55, 63] have demonstrated similar or slightly better accuracy on test data after training a GP model and using it to calibrate a physics-based model. Besides the obvious drawback of the increased computational time to

query the model in comparison to the nearly negligible time to query a GP, this method has an additional drawback of a needing to use Bayesian method to perform the calibration. These Bayesian methods, while becoming more common, can be costly and require some subject matter expertise to obtain convergence. The method implemented here uses a more naive approach. While each model has the same general material of 316L stainless steel, none of the models have been specifically calibrated to produce the same results. Thus, for a new material system, one could simply obtain data and/or model parameters from literature or utilize existing models to generate a new predictive model without the need to perform a new calibration.

There are a number of possible improvements that could be made to the NARGP to perform regression, such as the inclusion of simulation data from a multiphysics model, but one of the most notable improvements would be to change the NARGP from having multiple independent outputs to being a multivariate or multioutput model [50]. The process to convert the NARGP to a multivariate model would be relatively straightforward, though still beyond the scope of this work. Since the NARGP essentially only requires the iterative training of multiple GPs, rather a single multifidelity GP, the process to convert to a multivariate model would only entail converting each GP into a multivariate GP and training the NARGP as usual. Converting to a multivariate model could accomplish two objectives. First, it could potentially improve the regression model by leveraging information from both width and depth simultaneously while training. Secondly, by having a multivariate model, Bayesian methods could be used to impute the missing melt pool length data when it is unavailable. The imputation of the missing data would be possible without a multivariate model, but with a multivariate model, the coupling of the melt pool length, width, and depth at lower fidelities could greatly improve the imputation. In addition to the improvements discussed on the GP modeling methodology, additional improvements may be possible by including additional process parameter inputs. However, given the accuracy achieved using only power, velocity, and spot size, it is unlikely that additional parameters would dramatically improve the results.

Classification

The purpose of classification is to determine the material printability map, in terms of porosity defect type, based on the AM process parameters. Unlike regression, classification will always be performed on all available data. For the sake of this work, classification will be conducted in a number of ways for comparison purposes. First, it is performed as described by “Classification” section for each of the three defect classes, where each class is binary 1 for defect and 0 if that defect is not present. A second approach is to group

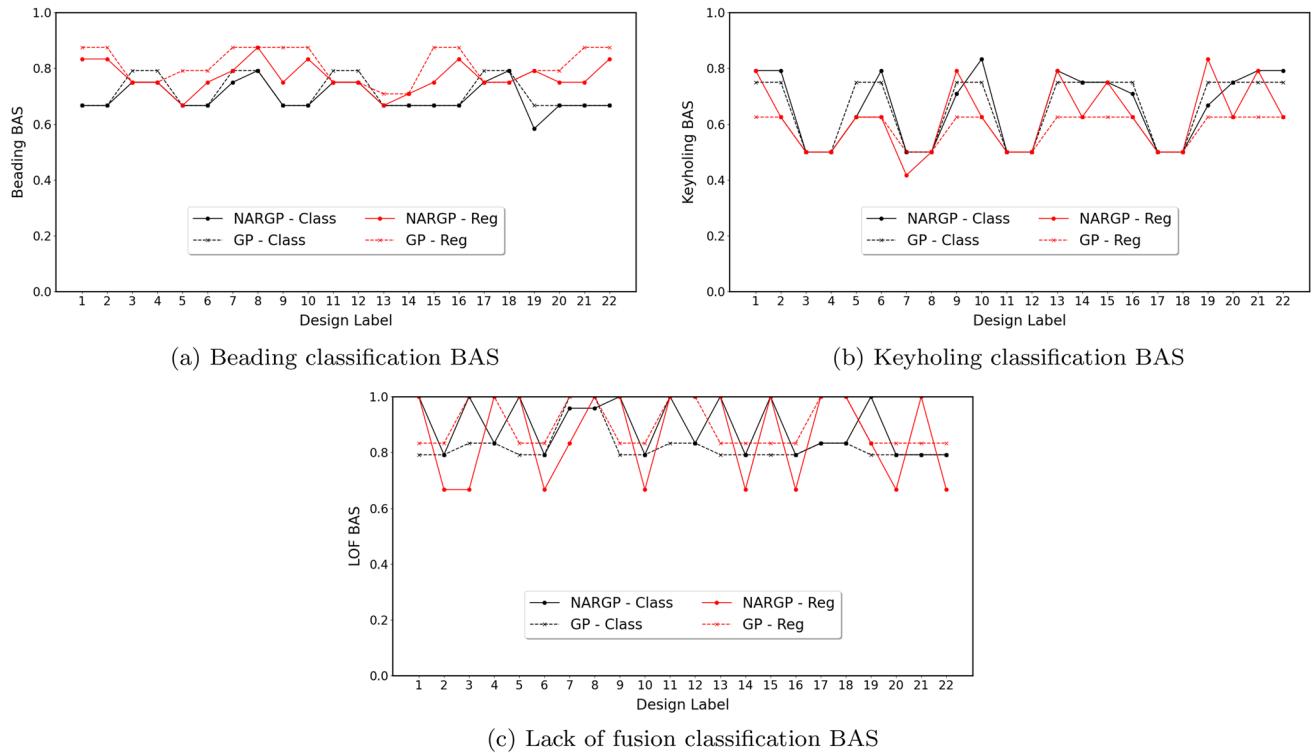


Fig. 7 Classification BAS for three defect types for the NARGP and GP classifiers and NARGP and GP regression-based classifiers

all the defects into a binary problem of defect (1) or no defect (0) if one or more of the defects is present. Finally, an approach similar to previous approaches [14, 17], where a trained regression model and empirical measures, as detailed above, are used to predict if a defect will be present and what that defect will be. As the beading metric requires melt pool length to approximate, any NARGP or GP that contains EXP will use the melt pool width and depth as normal but use melt pool length predictions based on the next lowest fidelity in the NARGP (*e.g.*, an NARGP trained with EXP, FE, and NEASM will have melt pool length NARGP trained only on FE and NEASM and GP trained only on FE while EXP would be included in NARGP and GP for width and depth training). To evaluate each of the classification models, a balanced accuracy score (BAS) is used. The BAS is a standard measure of classification accuracy except each class is weighted by the number of samples in that class. The BAS is used in this work since the number of defects in the test set for each defect type is relatively low (3 of each type over the 15 total points in the test set). Note that the BAS is 1 for a perfect classifier and is bounded between 0 and 1.

The classification results when predicting each defect individually are shown in Fig. 7. In the case of beading (Fig. 7a), a maximum BAS of 0.875 is achieved and is obtained using the regression-based NARGP and GP. The NARGP is only able to gain this level of accuracy using only the NEASM and ET models with a linear correlation.

The GP on the other hand obtains this accuracy for a model trained on the NEASM. A confusion matrix for this case is shown in Table 2. Interestingly, the NARGP and GP classifiers can only obtain approximately 0.8 BAS. This is obtained for a number of combinations but most notably, none of the combinations involve using experiments. For keyholing (Fig. 7b), a maximum BAS of 0.83 is obtained with the NARGP classifier as well as regression-based classifier. For the classifier, this BAS is achieved using EXP, FE, and ET runs with a linear correlation and for the regression, it is achieved with EXP and NEASM runs using a nonlinear correlation. The confusion matrices are identical for the two best classifiers and one is show in Table 3. Interestingly, this maximum BAS regression case does not correspond to the regression results that best predict melt pool width and depth. In the case of lack of fusion (Fig. 7c), there a number of models that result in perfect predictions from all classifiers. The most notable result is with the NEASM and

Table 2 Confusion matrix for the best beading classification (GP regression-based classifier trained using NEASM)

	No predicted defect	Predicted defect
No actual defect	9	3
Actual defect	0	3

Table 3 Confusion matrix for the best keyholing classification (NARGP regression-based classifier trained using EXP and NEASM with a nonlinear correlation)

	No predicted defect	Predicted defect
No actual defect	12	0
Actual defect	1	2

ET models combined or simply the NEASM, regardless of which method is used, can result in the highest BAS. Since the best cases result in perfect predictions, confusion matrices are not shown as they are simple matrices with zero as the off-diagonal components.

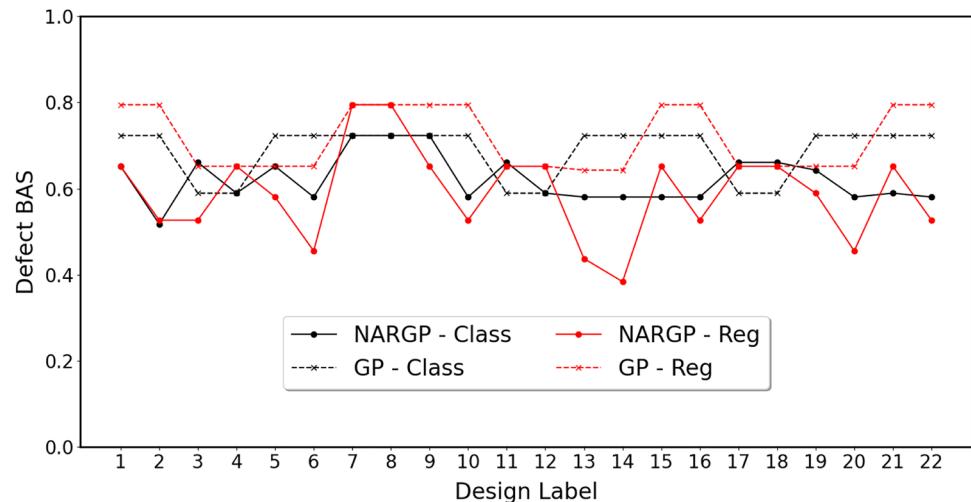
Again, note that the previous classifiers were tested on a set of 15 points where only 3 points for each of the defects were present in the 15. This results in highly skewed classes and makes determining the true model accuracy harder. By combining all defects into a single class, the skew becomes much less noticeable as there are seven defects of the 15 points. Combining the classes also reduces training time as only a single model needs to be trained for the classifier. For the regression-based classifier, three models are still trained since all melt pool dimensions are still needed to make classifications before being combined into a single class. The results of the binary defect classification are shown in Fig. 8. The results show a highest possible BAS of around 0.8 for the regression-based classifiers. These occur for GPs trained on experiment data and NARGP and GPs trained on NEASM and ET and NEASM, as was seen before. As before, each of these best models results in an identical confusion matrix and one example is shown in Table 4.

A few interesting points can be noted about both the combined and individual defect classifiers. First, it appears that in many cases, the regression-based classifier (*i.e.*, empirical classification measures from melt pool dimensions) shows similar or better performance than the regular classifiers.

Table 4 Confusion matrix for the best classification using the grouped binary classifier (GP regression-based classifier trained using EXP)

	No predicted defect	Predicted defect
No actual defect	5	2
Actual defect	1	7

Furthermore, the NARGP classifier does not appear to outperform a standard GP classifier in most cases. This is in contrast to the regression where the NARGP could greatly outperform the standard GP. One possible explanation for the lack of performance of the NARGP classifier is that the BAS is only indicating the mean performance of the classifier and not the full probability space. In testing of the NARGP on a pedagogical example, it was noted that the NARGP classifier produced fewer correct classifications but the overall probability map was closer to the true boundary used to generate the data. It is possible that this is again occurring here, however, without a concrete means to assess this, the BAS and similar metrics are the only option to evaluate a model. Comparing the two approaches to classification, grouped versus individual, there seems to be little impact on the BAS results. In the individual cases, the NARGP and GP classifiers both perform around 0.8, which is close to the highest level of achievable accuracy in the grouped cases. However, this conclusion is somewhat superfluous as the regression-based classifier, using empirical measures derived from melt pool dimensions, appears to outperform the GP classifier, and the regression approach only makes predictions without considering individual versus grouped classes during training. An interesting point that was mentioned earlier is that a GP classifier trained on the NEASM as the highest or only fidelity tends to produce the best or nearly the best results in all cases except individual keyholing predictions, where the experiments produce the

Fig. 8 Grouped binary classifier BAS for NARGP and GP classifiers and NARGP and GP regression-based classifiers

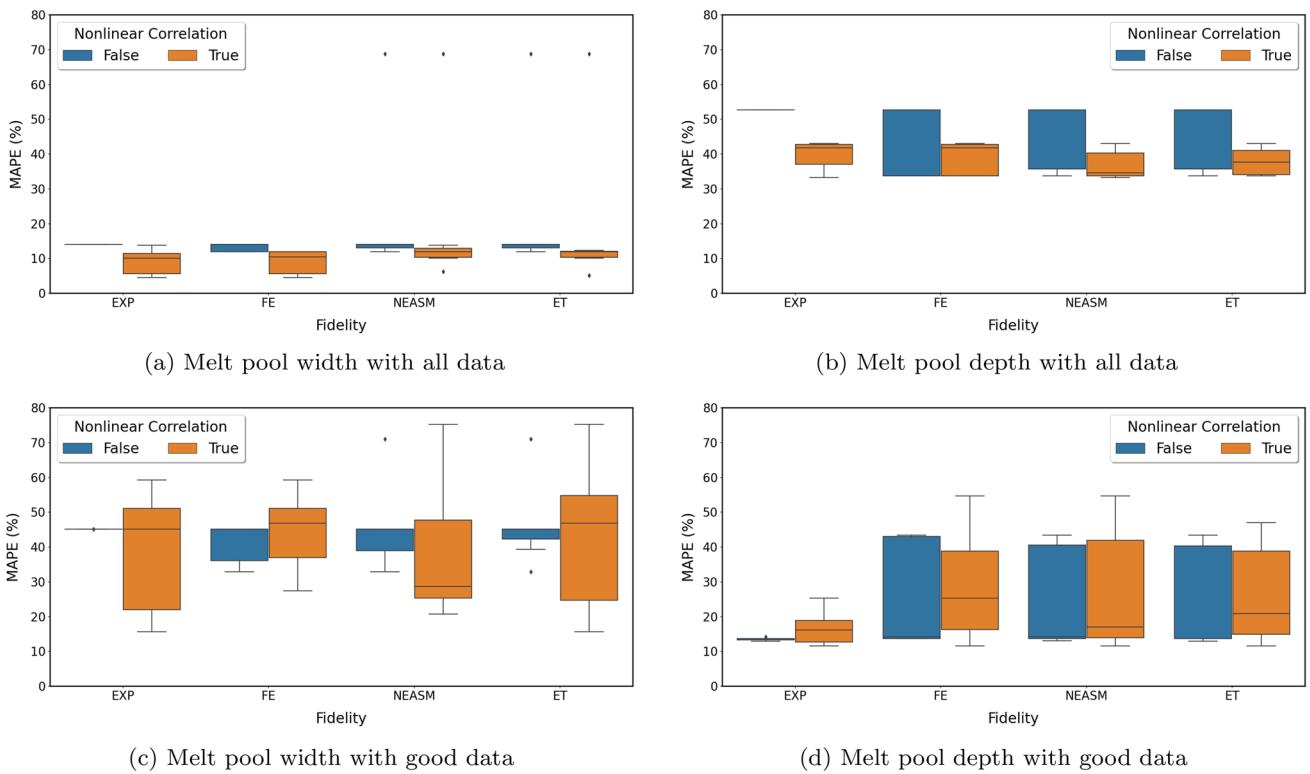


Fig. 9 Box plots showing MAPE ranges for NARGP models which included the given fidelity, both, with and without a nonlinear correlation

best model likely because the experiments are the only fidelity with the necessary physics to capture keyholing.

As with the regression there are some improvement that could be made to the classification. The most notable improvement would be to change to a multiclass classifier. This would nullify the need to decide between a binary defect/no-defect class or training individual classifiers. Much like the regression, changing to a multiclass classifier would be as simple as replacing the GP classifier implemented in this work with a new variant. The difficulty in changing to a multiclass classifier in the NARGP context is that the most common method to perform multiclass GP classification is via Bayesian methods, which become impractical when training multiple models sequentially. However, some recent works have begun to develop and implement non-Bayesian approaches to multiclass GP classification, which could be implemented in the NARGP framework. As with the NARGP regression problem, further improvements may be realized in the classification if additional input parameters are used that may help determine whether a defect will be induced or not. However, in the absence of additional data, especially from experiments, it is uncertain whether adding more inputs will improve the classification models or not.

Process Model Importance

In the previous section, each of the 22 trained NARGPs was examined and compared individually to the test set of experiments. Now, the NARGP results will be used to determine what fidelities included in the results tend to result in a better performing model. This is done by grouping the MAPE and BAS for each NARGP that includes each fidelity. This is further split by separating the linear versus nonlinear correlation models. Box plots of the data generated by this process are shown in Fig. 9 for the melt pool width and depth MAPE (for models with all data used for training and only good data used) and in Fig. 10 for the grouped BAS from the NARGP classifier and NARGP regression-based classification. For brevity, only the grouped classification results are examined, and R^2 is omitted since it generally agrees with the MAPE metric.

In the case of the regression results for melt pool width, it can be noticed that a nonlinear correlation shows a wider spread of MAPE values, but tends to produce a better result in both cases of training with all data or only good data. In the case of the linear correlation, the melt pool width MAPE does not tend to change much as different fidelities are included in the NARGP, whereas for the nonlinear correlation, the MAPE does tend to slightly decrease as higher fidelity data is added to the NARGP as one would

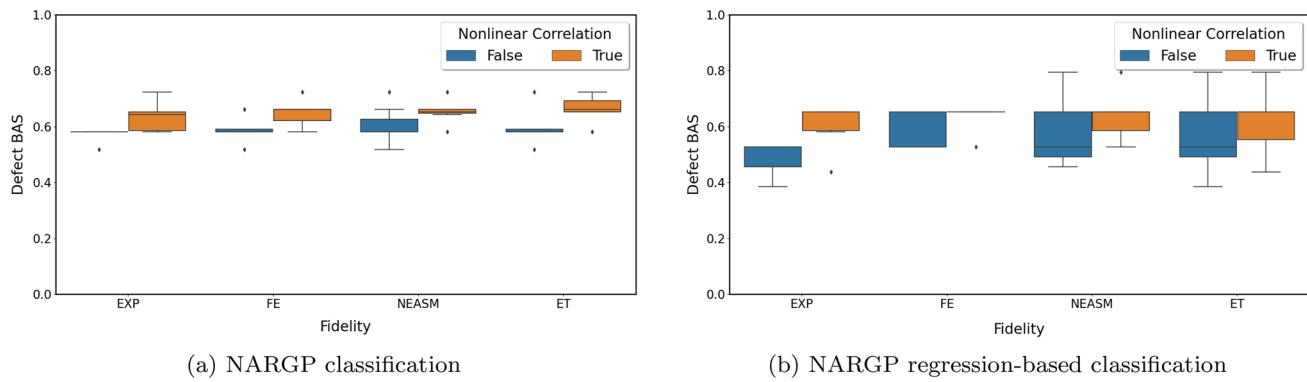


Fig. 10 Grouped defect BAS box plots for fidelities included in the NARGP model for models with and without nonlinear correlations

expect. The results on melt pool width with all data versus good data also reinforce the conclusions above that a low MAPE can be achieved with a model trained on all data. With the exception of a couple outliers, the models trained on all data also show a lower variance than the models trained on good data only.

For melt pool depth, the conclusions from the previous section still hold, in that NARGP models trained on good data tend to produce lower error rates than a model trained on all data. Furthermore, the models trained using experiments result in the lowest MAPE values with the lowest variance. In the case of melt pool depth, the variance of linear correlation versus nonlinear correlation, as well as variance changes with fidelity, do not appear to change significantly until experiments are included. This, as stated, reduces the variance and tends to give the best results. Unlike melt pool width, the melt pool depth does not seem to improve significantly with higher fidelity data. The best model is any model multifidelity model that includes experiments. A similar conclusion can be made about melt pool width.

Unlike regression, the classification shows consistent BAS values regardless of NARGP fidelities or correlation used. In fact, there is a slight upward trend in the BAS for the NARGP classifier when lower fidelities are included. A similar trend can be noted for the regression-based classifier, although the trend is not as prevalent. The trends shown in the box plots offer additional insight into the previous conclusion about the regression-based classification outperforming the standard classification. The NARGP classification, as detailed in this work, does tend to perform as well or better than the regression-based version. However, the highest BAS obtained is with the regression-based models. The classification method presented in this work also shows a lower variance, which could be a desirable feature that makes selection of specific fidelity data sets less important.

Model Selection

Having several fidelities of data available to make the best model is obviously ideal, but training all combinations of models and generating data from so many sources is generally not practical. Here, some best practices are discussed that result in generating data from the fewest fidelities that result in the best trained models. This is done for regression only models, classification only models, and then models for classification and regression.

For regression, two models will likely need to be trained regardless; one for melt pool width on all available data and another for melt pool depth where the data has been screened *a priori* to select only data which does not produce defects. While training two models is not ideal, the benefits could be worthwhile since a melt pool width prediction of <5% and depth prediction of 11% is possible. Additionally, the NARGP is implemented with a maximum likelihood approach so training of a model is relatively fast. The drawback to achieve the level of accuracy shown is that at least three fidelities of data would be required, namely EXP, FE, and NEASM. This could be a significant cost to acquire data from those sources and may not even be possible given that enriched analytical solutions are not widely available. Alternatively, a model using only EXP and ET data can be trained which results in a greatly reduced data acquisition cost with only a slight accuracy penalty, achieving 12.3% width and 13.6% error rates for the melt pool width and depth, respectively. While this level of error is somewhat high, it could serve as a good first order approximation and does not require specific calibration of the ET parameters using Bayesian methods.

In the case of classification, recommendations are much more difficult to make as clear trends are not seen as has been demonstrated in the previous sections. Using the NEASM model for any of the four classification methods shown (NARGP and GP classification and regression-based classification) appears to result in the best classifier for both

individual and grouped classification except for keyholing predictions. Only models containing experiments are able to accurately predict keyholing to some extent. For the grouped classification, similar conclusions can be drawn. A GP regression-based classification with the NEASM or EXP provides the highest BAS around 0.8. Interestingly, when combined in a multifidelity model, the result does not improve and tends to get slightly worse at 0.72 BAS.

Taking everything together, an NARGP model containing EXP, NEASM, and ET seems to be able to provide the best predictive model. The experiments improve keyholing and melt pool width and depth predictions, while the NEASM and ET have the purpose of predicting beading, lack of fusion, and/or grouped classification and add additional information to supplement the limited experiment data. While the resulting model would not be the best possible model, it would balance the data generation and model creation processes. Additionally, it would only require training a single NARGP regression model. While using the regression-based classification resulted in the best overall results, using the NARGP classification does have the potential to produce equivalent results to the regression-based version with less dependence on which fidelities are selected as was shown in the previous section.

Conclusions

This work has presented an analysis and discussion on predicting and classifying AM melt pools using multifidelity GP surrogates. The work considered four common sources of information at different fidelities, from analytical to experimental, and combined that information in a NARGP regression and novel classification models. The models were trained on a representative set of data and compared to a test set of experiments. All combination of fidelities were trained and compared to a standard GP. The NARGP regression demonstrated superior results over standard GP regression and require no model calibration *i.e.*, Bayesian methods. Furthermore, certain combinations of models, namely the experiments and Eagar–Tsai model, were able to produce an accurate model which balanced data generation cost with attainable model accuracy. Depending on the fidelity of information available, a model was able to be produced, which matched or exceeded the performance of existing GP modeling and calibration approaches. For classification, results using both grouped and ungrouped classes were examined along with regression-based classification using empirical measures and how that compared to using binary classification methods. The results showed that for ungrouped classification, experiments were necessary to accurately predict keyholing, as one would expect, but using only the NEASM or a multifidelity model with

the NEASM was sufficient to predict beading and lack of fusion. In the grouped classification setting, similar results were seen. In both cases, the regression-based classification was shown to produce a singular best model for classifying printability, but typical classification methods were able to produce more consistent results that were less dependent on the fidelities included in the model. While the NARGP classification approaches were able to produce some models that exceeded the performance of a standard GP, the results were not consistent. It was postulated that while the NARGP did not improve significantly on the classification accuracy for the available test set, it likely improved the predicted probability space.

Acknowledgements RS, AR, AB, AI, and JM acknowledge partial support for this work by the Office of Naval Research through the Naval Research Laboratory's (NRL) core funding. RS acknowledges partial support for this work by the NRL Edison Graduate Memorial Training Program.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. The White House Office of Science and Technology Policy (2022) Fast track action subcommittee on critical emerging technologies: critical and emerging technologies list update. <http://www.whitehouse.gov/ostp>
2. Herzog D, Seyda V, Wycisk E, Emmelmann C (2016) Additive manufacturing of metals. Acta Materialia 117:371–392. <https://doi.org/10.1016/j.actamat.2016.07.019>
3. The White House (2022) Fact sheet: Biden administration celebrates launch of am forward and calls on congress to pass bipartisan innovation act. <https://www.whitehouse.gov/briefing-room/statements-releases/2022/05/06/fact-sheet-biden-administration-celebrates-launch-of-am-forward-and-calls-on-congress-to-pass-bipartisan-innovation-act/>
4. Exec. order no. 14,017, 86 c.f.r. 11849. (2021)
5. U.S. Department of Defense (2022) Securing defense-critical supply chains. Tech. rep.
6. U.S. Department of Energy (2022) America's strategy to secure the supply chain for a robust clean energy transition. Tech. rep.
7. Under Secretary of Defense for Research and Engineering (2022) USD(R&E) Technology Vision for an Era of Competition. Tech. rep, US Department of Defense
8. DebRoy T, Wei H, Zuback J, Mukherjee T, Elmer J, Milewski J, Beese A, Wilson-Heid A, De A, Zhang W (2018) Additive manufacturing of metallic components - process, structure and properties. Progr Mater Sci 92:112–224. <https://doi.org/10.1016/j.pmatsci.2017.10.001>
9. Gao W, Zhang Y, Ramanujan D, Ramanan K, Chen Y, Williams CB, Wang CC, Shin YC, Zhang S, Zavattieri PD (2015) The status, challenges, and future of additive manufacturing in engineering. Comput Aided Des 69:65–89. <https://doi.org/10.1016/j.cad.2015.04.001>

41. Ganeriwala R, Zohdi TI (2016) A coupled discrete element-finite difference model of selective laser sintering. *Granular Matter* 18(2):1–15. <https://doi.org/10.1007/s10035-016-0626-0>
42. Khairallah SA, Anderson AT, Rubenchik AM, King WE (2016) Laser powder-bed fusion additive manufacturing: physics of complex melt flow and formation mechanisms of pores, spatter, and denudation zones. *Acta Materialia* 108:36–45. <https://doi.org/10.1016/j.actamat.2016.02.014>
43. Steuben JC, Iliopoulos AP, Michopoulos JG (2016) Discrete element modeling of particle-based additive manufacturing processes. *Comput Methods Appl Mech Eng* 305:537–561. <https://doi.org/10.1016/j.cma.2016.02.023>
44. Panwisawas C, Qiu C, Anderson MJ, Sovani Y, Turner RP, Attallah MM, Brooks JW, Basoalto HC (2017) Mesoscale modelling of selective laser melting: thermal fluid dynamics and microstructural evolution. *Comput Mater Sci* 126:479–490. <https://doi.org/10.1016/j.commatsci.2016.10.011>
45. Pei W, Zhengying W, Zhen C, Junfeng L, Shuzhe Z, Jun D (2017) Numerical simulation and parametric analysis of selective laser melting process of AlSi10Mg powder. *Appl Phys A Mater Sci Process* 123(8):1–15. <https://doi.org/10.1007/s00339-017-1143-7>
46. Xia M, Gu D, Yu G, Dai D, Chen H, Shi Q (2017) Porosity evolution and its thermodynamic mechanism of randomly packed powder-bed during selective laser melting of Inconel 718 alloy. *Int J Mach Tools Manuf* 116:96–106. <https://doi.org/10.1016/j.ijmachtools.2017.01.005>
47. Moges T, Ameta G, Witherell P (2019) A review of model inaccuracy and parameter uncertainty in laser powder bed fusion models and simulations. *J Manuf Sci Eng* 141(4):1. <https://doi.org/10.1115/1.4042789>
48. Tapia G, Elwany AH, Sang H (2016) Prediction of porosity in metal-based additive manufacturing using spatial Gaussian process models. *Addit Manuf* 12:282–290. <https://doi.org/10.1016/j.addma.2016.05.009>
49. Tapia G, Johnson L, Franco B, Karayagiz K, Ma J, Arroyave R, Karaman I, Elwany A (2017) Bayesian calibration and uncertainty quantification for a physics-based precipitation model of nickel-titanium shape-memory alloys. *J Manuf Sci Eng* 139(7):071002. <https://doi.org/10.1115/1.4035898>
50. Mahmoudi M, Tapia G, Karayagiz K, Franco B, Ma J, Arroyave R, Karaman I, Elwany A (2018) Multivariate calibration and experimental validation of a 3d finite element thermal model for laser powder bed fusion metal additive manufacturing. *Integr Mater Manuf Innov* 7(3):116–135. <https://doi.org/10.1007/s40192-018-0113-z>
51. Tapia G, Khairallah S, Matthews M, King WE, Elwany A (2018) Gaussian process-based surrogate modeling framework for process planning in laser powder-bed fusion additive manufacturing of 316L stainless steel. *Int J Adv Manuf Technol* 94(9–12):3591–3603. <https://doi.org/10.1007/s00170-017-1045-z>
52. Tapia G, King W, Johnson L, Arroyave R, Karaman I, Elwany A (2018) Uncertainty propagation analysis of computational models in laser powder bed fusion additive manufacturing using polynomial chaos expansions. *J Manuf Sci Eng*. <https://doi.org/10.1115/1.4041179>
53. Nath P, Hu Z, Mahadevan S (2019) Uncertainty quantification of grain morphology in laser direct metal deposition. *Model Simul Mater Sci Eng*. <https://doi.org/10.1088/1361-651X/ab1676>
54. Wang Z, Liu P, Ji Y, Mahadevan S, Horstemeyer MF, Hu Z, Chen L, Chen LQ (2019) Uncertainty quantification in metallic additive manufacturing through physics-informed data-driven modeling. *JOM* 71(8):2625–2634. <https://doi.org/10.1007/s11837-019-03555-z>
55. Honarmandi P, Arróyave R (2020) Uncertainty quantification and propagation in computational materials science and simulation-assisted materials design. *Integr Mater Manuf Innov* 9(1):103–143. <https://doi.org/10.1007/s40192-020-00168-2>
56. Ye J, Mahmoudi M, Karayagiz K, Johnson L, Seede R, Karaman I, Arroyave R, Elwany A (2022) Bayesian calibration of multiple coupled simulation models for metal additive manufacturing: a bayesian network approach. *ASCE-ASME J Risk and Uncert in Engrg Sys Part B Mech Engrg* 8(1):1–12. <https://doi.org/10.1115/1.4052270>
57. Kennedy M (2000) Predicting the output from a complex computer code when fast approximations are available. *Biometrika* 87(1):1–13. <https://doi.org/10.1093/biomet/87.1.1>
58. Higdon D, Kennedy M, Cavendish JC, Cafeo JA, Ryne RD (2004) Combining field data and computer simulations for calibration and prediction. *SIAM J Sci Comput* 26(2):448–466. <https://doi.org/10.1137/S1064827503426693>
59. Higdon D, Gattiker J, Williams B, Rightley M (2008) Computer model calibration using high-dimensional output. *J Am Stat Assoc* 103(482):570–583. [https://doi.org/10.1198/0162145070000000888](https://doi.org/10.1198/016214507000000888)
60. COMSOL AB: Comsol multiphysics® v5.6. www.comsol.com. Stockholm, Sweden
61. Seede R, Shoukr D, Zhang B, Whitt A, Gibbons S, Flater P, Elwany A, Arroyave R, Karaman I (2020) An ultra-high strength martensitic steel fabricated using selective laser melting additive manufacturing: densification, microstructure, and mechanical properties. *Acta Materialia* 186:199–214. <https://doi.org/10.1016/j.actamat.2019.12.037>
62. Tapia G, King WE, Arroyave R, Johnson L, Karaman I, Elwany A (2018) Validation of a laser-based powder bed fusion thermal model via uncertainty propagation and generalized polynomial chaos expansions. *J Manuf Sci Eng*. <https://doi.org/10.1115/1.4041179>
63. Honarmandi P, Seede R, Xue L, Shoukr D, Morcos P, Zhang B, Zhang C, Elwany A, Karaman I, Arroyave R (2021) A rigorous test and improvement of the Eagar-Tsai model for melt pool characteristics in laser powder bed fusion additive manufacturing. *Addit Manuf* 47(August):102300. <https://doi.org/10.1016/j.addma.2021.102300>
64. Michopoulos JG, Steuben JC, Birnbaum AJ, Iliopoulos AP, Aroh J, Rollett AD, Gould B (2020) Morphological analysis of 316L laser powder bed fusion melt-pool via the enriched analytical solution method. In: Volume 9: 40th Computers and Information in Engineering Conference (CIE), pp. 1–14. American Society of Mechanical Engineers . [https://asmedigitalcollection.asme.org/IDETC-CIE/proceedings/IDETC-CIE2020/83983/Virtual, Online/1090045](https://doi.org/10.1115/DETC2020-22455)
65. The Mathworks Inc.: Matlab® v2021b. www.mathworks.com. Natick, Massachusetts , USA
66. Rawlings ALK, Birnbaum AJ, Michopoulos JG, Steuben JC, Iliopoulos AP, Ryou H (2020) Simulation informed effects of solidification rate on 316L single tracks produced by selective laser melting. In: Volume 9: 40th Computers and Information in Engineering Conference (CIE), February 2021. American Society of Mechanical Engineers . [https://asmedigitalcollection.asme.org/IDETC-CIE/proceedings/IDETC-CIE2020/83983/Virtual, Online/1090040](https://doi.org/10.1115/DETC2020-22451)
67. Le Gratiet L, Garnier J (2014) Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *Int J Uncertain Quantif* 4(5):365–386. <https://doi.org/10.1615/Int.J.UncertaintyQuantification.2014006914>
68. Le Gratiet L, Cannamela C (2015) Cokriging-based sequential design strategies using fast cross-validation techniques for multi-fidelity computer codes. *Technometrics* 57(3):418–427. <https://doi.org/10.1080/00401706.2014.928233>
69. Perdikaris P, Raissi M, Damianou A, Lawrence ND, Karniadakis GE (2017) Nonlinear information fusion algorithms for

- data-efficient multi-fidelity modelling. Proc Royal Soc A Math Phys Eng Sci. <https://doi.org/10.1098/rspa.2016.0751>
- 70. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, Cambridge
 - 71. Williams C, Barber D (1998) Bayesian classification with Gaussian processes. IEEE Trans Patt Anal Mach Intell 20(12):1342–1351. <https://doi.org/10.1109/34.735807>
 - 72. Sahli Costabal F, Perdikaris P, Kuhl E, Hurtado DE (2019) Multi-fidelity classification using Gaussian processes: accelerating the prediction of large-scale computational models. Comput Methods Appl Mech Eng 357:112602. <https://doi.org/10.1016/j.cma.2019.112602>
 - 73. Klyuchnikov N, Burnaev E (2020) Gaussian process classification for variable fidelity data. Neurocomputing 397:345–355. <https://doi.org/10.1016/j.neucom.2019.10.111>
 - 74. Qian PZ (2009) Nested Latin hypercube designs. Biometrika 96(4):957–970. <https://doi.org/10.1093/biomet/asp045>
 - 75. He X, Qian PZ (2011) Nested orthogonal array-based Latin hypercube designs. Biometrika 98(3):721–731. <https://doi.org/10.1093/biomet/asr028>
 - 76. Xu J, Duan X, Wang Z, Yan L (2018) A general construction for nested Latin hypercube designs. Stat Probab Lett 134(11771450):134–140. <https://doi.org/10.1016/j.spl.2017.10.022>
 - 77. Dash S, Mandal BN, Parsad R (2020) On the construction of nested orthogonal Latin hypercube designs. Metrika 83(3):347–353. <https://doi.org/10.1007/s00184-019-00721-w>
 - 78. Paleyes A, Pullin M, Mahsereci M, Lawrence N, González J (2019) Emulation of physical processes with Emukit. In: Second Workshop on Machine Learning and the Physical Sciences, NeurIPS
 - 79. Kim S, Kim H (2016) A new metric of absolute percentage error for intermittent demand forecasts. Int J Forecast 32(3):669–679. <https://doi.org/10.1016/j.ijforecast.2015.12.003>