Final Report:

Supervised linear regression machine learning model to predict the primary dendritic arm spacing

Sepideh Kavousi

Problem Statement

During the solidification of metallic systems, the microstructure grows in treelike dendritic shape and the morphology of dendrites has a strong effect on the mechanical properties. Spacing between the dendrite arms provides information on segregation patterns and distribution of precipitates between the dendritic arms, which influences the mechanical properties. There are two well-known theoretical models that relate the primary dendritic arm spacing (PDAS) to solidification conditions, namely pulling velocity (V) and temperature gradient (G), and material properties like liquid diffusivity (D_l), the freezing range (ΔT_0), Gibbs-Thomson coefficient (Γ), and equilibrium partition coefficient (k_e). The two models, named Hunt-Burden (HB) [1] and Kurz-Fisher (KF) [2], are different in the way the material properties are accounted:

$PDAS = 2.83(k_e \Gamma \Delta T_0 D_l)^{0.25} G^{-0.5} V^{-0.25}$ HB model

$PDAS = 4.3(\Gamma \Delta T_0 D_l / k_e)^{0.25} G^{-0.5} V^{-0.25}$. KF model

Despite the previous attempts, there are still important issues to develop a predictive model of PDAS for a wide range of solidification rates. In rapid solidification cases, such as in laser additive manufacturing of materials, the analytical relations and models for PDAS have significant deviations from computational predications and experimental data for the same solidification condition [3]. In this project we want to develop more reliable analytical models for prediction of PDAS in a wide range of solidification rates. For this, a supervised linear regression machine learning model will be used to produce a new relation between PDAS and the solidification rate (pulling velocity and temperature gradient in the model) that works for different material systems.

Data Generating

PDAS data for the PDAS versus solidification rate and for different materials systems is generated for 7 alloy systems, with various solidification velocity range (from 10^{-4} m/s to 0.12 m/s), temperature gradient range (from 10^6 K/m to 10^7 K/m) of 10 different alloys to generate enough data points for machine learning analysis. Obtaining quantitative prediction of the characteristics of solidification process with phase field simulation approach requires additional knowledge of crystal melt properties, such as anisotropic crystal-melt interfacial free energy and anisotropic kinetic coefficient. A major limiting factor is the scarcity of available experimental data pertaining to crystal-melt interfacial properties. Integrated atomistic and phase-field simulations will be performed in order to investigate the solidification of binary alloys.

In a separate study I have performed simulation study to obtain the computational data on PDAS for 7 different alloy systems (250 data points totally), for a wide velocity and temperature gradient range (given in table 1). We have also calculated all the material properties needed for each alloy systems ($k_e, \Gamma, \Delta T_0, D_l$).

Table 1- Summary of the computational PDAS and the range of velocity (V), temperature gradient (G), and the number of datapoints for each alloy system.

Alloy	V (m/s)	G (K/m)	Number of data points
Ti-3.4 at% Ni	0.01-0.09	10^7, 8×10^6, 7×10^6, 6×10^6, 5×10^6	45
Ti-7.1 at% Ni	0.01-0.09	10^7, 8×10^6, 7×10^6, 6×10^6, 5×10^6	42

Ti-10.4 at% Ni	0.0001-0.02	$10^{7}, 5\times10^{6}$	22
Mg-9 at% Al	0.005-0.12	$5\times10^{6}, 3\times10^{6}, 1\times10^{6}$	21
Al-6 at% Cu	0.001-0.06	$10^{7}, 8\times10^{6}, 7\times10^{6}, 6\times10^{6}, 5\times10^{6}$	47
Al-8 at% Cu	0.01-0.08	$10^{7}, 8\times10^{6}, 7\times10^{6}, 6\times10^{6}, 5\times10^{6}$	43
Al-10 at% Cu	0.008-0.005	$10^{7}, 8\times10^{6}, 7\times10^{6}, 6\times10^{6}$	30

Data Wrangling

It is important to obtain a clear visual view of the correlation between PDAS and the independent variables (material properties, V, and G). Visualization in this section helped us make sure we do not have any outliers in the PDAS data. The expected descending correlation between PDAS and V is more clear than the correlation between PDAs and G. However, as we have said previously, we need to check the logarithmic form of PDAS. The results do not present a specific trend in for G , mat_HB, mat_Kf in comparison with the

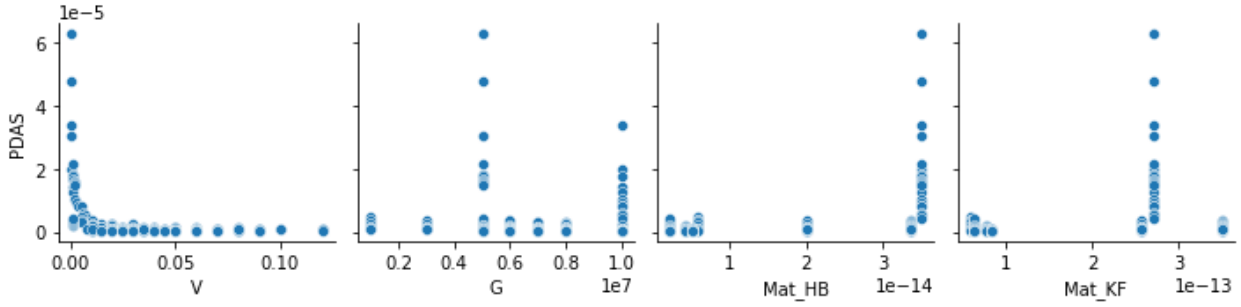

Figure 1 Variations of PDAS with V, G, material Properties.

In order to use a regression model, we will take the logarithm from both sides of the PDAS equation for both HB and KF models, while keeping the coefficients on ln(V), ln(G), ln(mat_HB), and ln(mat_KF) as unknown in the linear regression models.

$$\ln(PDAS) = \ln(\mathrm{A}) + \alpha \times \ln(V) + \beta \times \ln(G) + \gamma \times \ln(mat\_HB) \qquad mat_{HB} = k_e \Gamma \Delta T_0 D_L$$

$$\ln(PDAS) = \ln(\mathrm{A}) + \alpha \times \ln(V) + \beta \times \ln(G) + \gamma \times \ln(mat\_KF) \qquad mat_{KF} = \Gamma \Delta T_0 \, D_L / k_e$$

Then the data from computational PDAS will input into the linear regression models to obtain A, α, β, and γ.

We find a strong correlation (negative) between PDAS_ln and V_ln, also PDAS_ln and G_ln. both HB and KF material properties also have strong but positive effects on PDAS_ln. This contradicts our original models (HB and KF) which showed G playing a more dominant role on the PDAS calculations.

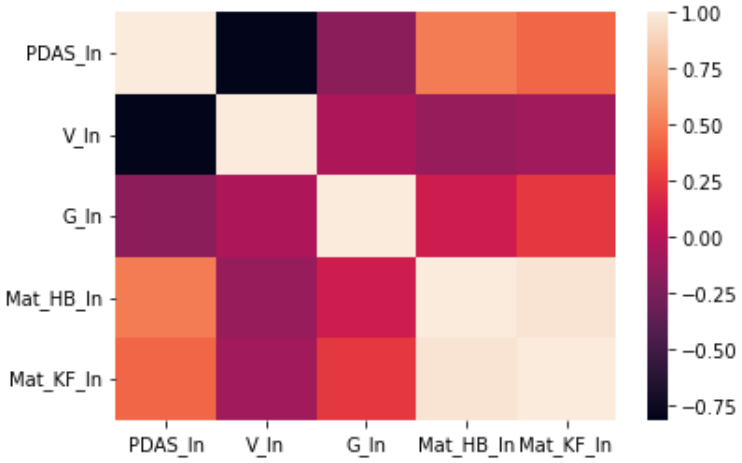

Figure 2 Heatmap graph based on Pearson correlation.

Modeling

Parametric ML models- Linear Regression Models

We used linear regression models using velocity, temperature gradient and material properties as the independent variables and the PDAS as the dependent variable. Meanwhile, instead of just one time dividing the data to test and training dataset, we used the k-fold cross cross validation model. In this method, the data are randomly divided into k groups, or folds, of approximately equal size. The first fold is treated as a test set, and the model is fit on the remaining k − 1 folds. The error is then computed on the observations in the held-out fold. This procedure is repeated k times; each time, a different group of observations is treated as a test set. This process results in k estimates of the test error. Figure 3 shows the performance of the linear regression model based on the HB and KF material properties, with the mean squared error (MSE) of 0.176 and 0.173, respectively.

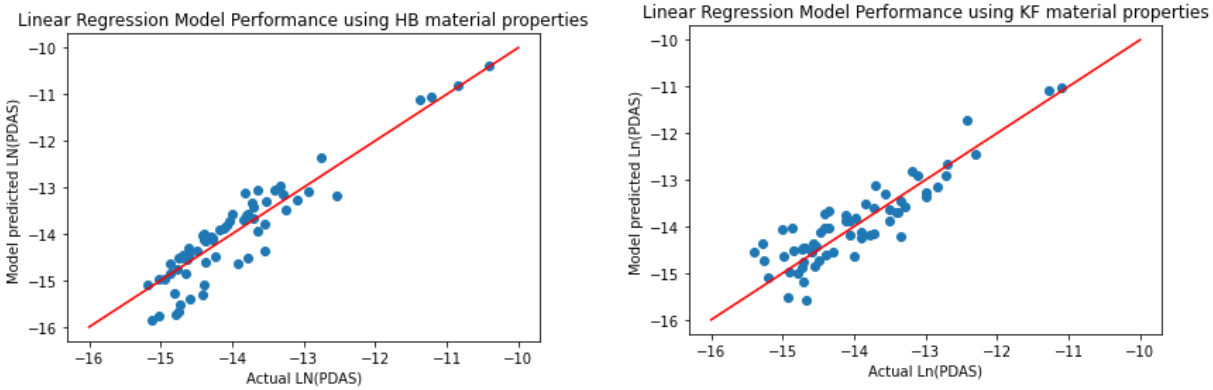

Figure 3 The actual Ln(PDAS) versus the model predictions using HB and KF material properties.

Nonparametric ML models

We also wanted to test different non parametric ML models. For this we performed hyperparameter tuning for k-nearest neighbor (KNN), random forest (RF), decision tree (DT) models. With the hyperparameter

branching, we found the optimum values for each model's parameters which minimized the MSE. Figure 4 presents the performance of both parametric and non-parametric ML models. In general, the non-parametric ML models show a much better performance in comparison to the parametric linear regression models. Even the difference between the regular linear regression models with Ridge or Lasso was not dramatically different. Among all the models the DT showed the best performance (0.0328).

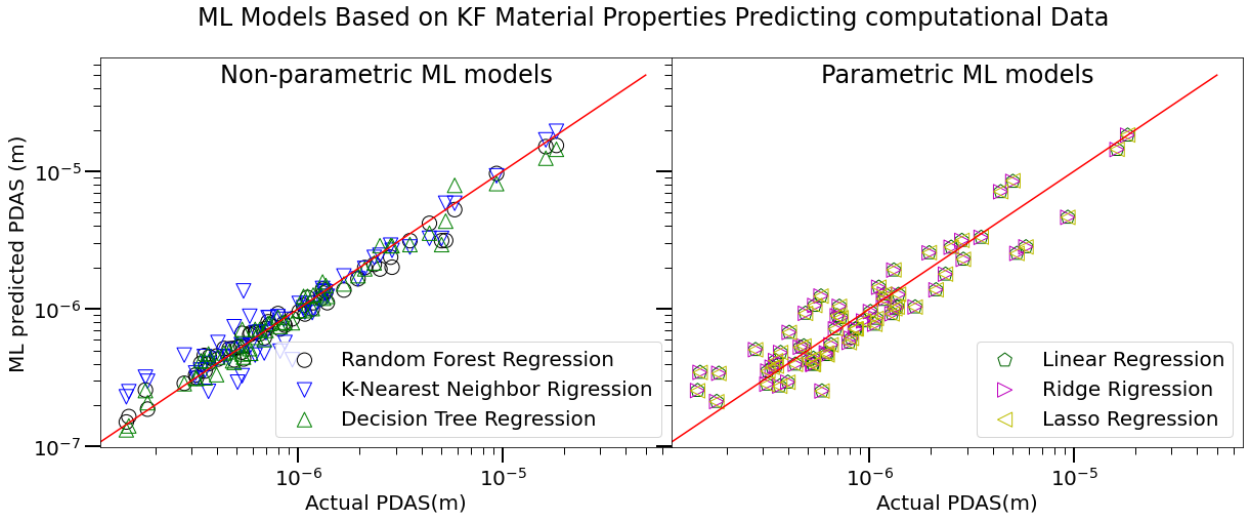

Figure 4 the comparison of parametric vs nonparametric ML models based on computational training dataset.

However, in order to validate our models, we checked tested the models using the experimental data and the results, as shown in figure 5, presented an opposite trend. The parametric linear regression models have a much better performance (MSE= 0.239) in comparison to RF (MSE=2.23), DT(MSE=1.68) and KNN (MSE=1.62) models. This can be due to the fact that i) Despite the inaccuracy of analytical model's coefficient, the general equation is correct, and that's the reason the regression models which follows the same equational forms fits bettwe with all the data ii) the dataset we had was very small so, the performance of the model for nonparametric models was high because the model was overfitted. Despite the performance of parametric models were

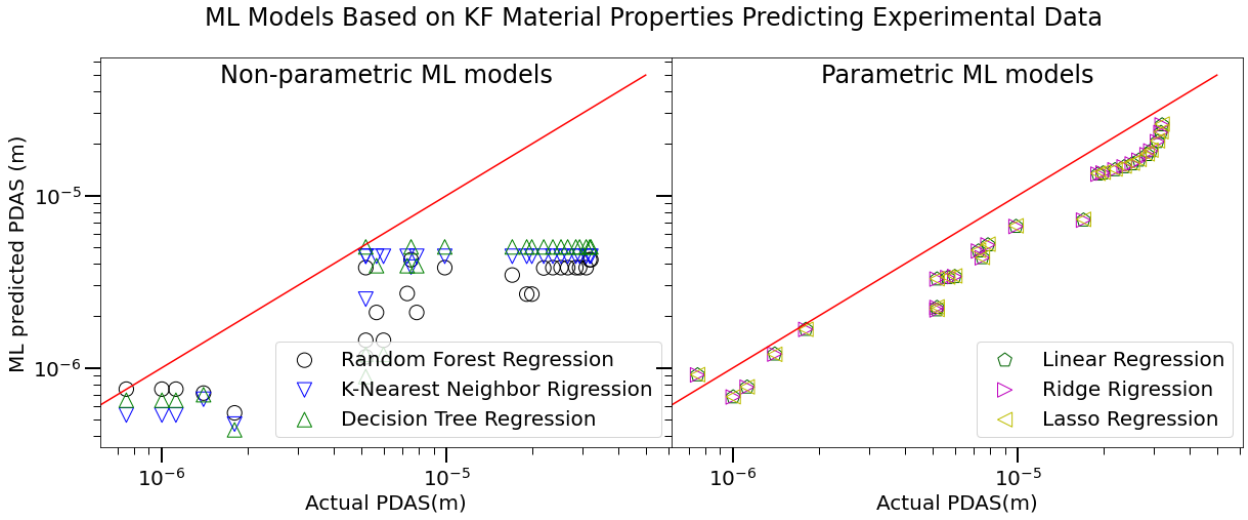

Figure 5 the comparison of parametric vs nonparametric ML models based on experimental training dataset.

Conclusion

1- We obtain a very important conclusion that in contract to analytical models, where temperature gradient plays much important role in the PDAS estimations, the velocity and temperature play almost equal roles in PDAS correlation!

2- In addition we compared different parametric and non parametric models. Despite the better performance of paranetric models on the computational data, when we compared them with the experimental predictions, they presented a much poorer performance and therefore, the parametric linear regression models works the best.

3- As we observed by the cross validation, the optimized alpha values for Ridge and Lasso regression models were very small. Therefore, we were already expecting the predictions of linear regression be close to Ridge and lasso models. As expected, the prediction of these models are very close to each other both for the experiment and computational data!

Suggestion

We suggest the fellow researchers to use the modified HB and KF models instead of the classical analytical ones:

$PDAS \propto V^{-0.7}\, G^{-0.68}\, (k_e \Gamma \Delta T_0 D_L)^{0.46}$ Modified HB model

$PDAS \propto V^{-0.72}\, G^{-0.84}\, (\Gamma \Delta T_0\, D_L / k_e)^{0.65}$ Modified KF model

References:

[1] M.H. Burden, J.D. Hunt, Cellular and dendritic growth. II, Journal of Crystal Growth, 22 (1974) 109-116.
[2] W. Kurz, D.J. Fisher, Dendrite growth at the limit of stability: tip radius and spacing, Acta Metallurgica, 29 (1981) 11-20.
[3] W. Xiao, S. Li, C. Wang, Y. Shi, J. Mazumder, H. Xing, L. Song, Multi-scale simulation of dendrite growth for direct energy deposition of nickel-based superalloys, Materials & Design, 164 (2019) 107553.