Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 126 (2024) 745–750

www.elsevier.com/locate/procedia

17th CIRP Conference on Intelligent Computation in Manufacturing Engineering (CIRP ICME '23)

Analysis of the relationship between training data volume and model quality for surrogate models in physical simulations

Tom Roeger[a,b,*], Ludwig Vogt[a], Tobias Friedrich[a], Johannes Schilp[a]

[a]*University of Augsburg, Universitätsstr. 2, 86159 Augsburg, Germany*

[b]*Grenzebach Maschinenbau GmbH, Albanusstraße 1, 86663 Asbach-Bäumenheim, Germany*

* Corresponding author. Tel.: +49 9069822089; *E-mail address:* tom.roeger@grenzebach.com

Abstract

Typically, the quality of AI models is highly dependent on the amount and quality of the available training data. While for many applications of AI several million training datasets are available in accessible databases and can be easily extended, the generation of training data for surrogate models of physics simulations is computationally demanding. To identify the necessary amount of training data for a sufficient good AI model, we are evaluating the performance of a surrogate model for a thermomechanical production process with different sizes of artificially generated training data.

Keywords: Quality; AI models; Training data; Surrogate models; Physics simulations; Computationally demanding; Thermomechanical production process; Model performance; Data size

1. Introduction

1.1 Background and Motivation

In recent years, the application of artificial intelligence (AI) and machine learning (ML) has developed rapidly in various fields. Particularly noteworthy are these developments in the area of modeling and simulation of physical processes. Surrogate models, also referred to as meta-models, have the potential to capture complex relationships between input and output parameters while significantly reducing computation time compared to traditional simulation methods. Thus, there would also be the possibility to use such methods for evaluation and recommendation in production environments, which were not previously possible due to their complexity or calculation time.

1.2 Objective and Research Questions

However, the quality of these models depends on the amount and quality of available training data. The generation of training data for surrogate models, which are obtained using physical models of the process steps, is costly and computationally intensive. Therefore, it is advantageous to know the required amount of training data in advance. For the reasons mentioned above, in this work, we have simulated a flat glass production process. We used the simulation to replicate the behavior of the process and to make predictions about the temperature profile during the manufacturing process in the product. Our intention was to reproduce this in a surrogate model, which is much faster in making a prediction and with a relatively short response time. This provides the opportunity to

10.1016/j.procir.2024.08.302

create a control system for complex production processes. For this purpose, we used the simulation to generate training data sets, and in the next step, we used a small number of training data sets to identify a suitable model architecture. Ultimately, we chose a feedforward model due to various advantages. Subsequently, we trained the selected model with different amounts of training data to establish a relationship between the amount of training data and model quality. We used the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) as quality criteria for the model. In this study, we aim to answer the following research questions: 1) How does the amount of training data influence the quality of surrogate models in physical simulations? 2) What is the optimal amount of training data required for a high-quality surrogate model?

2. Fundamentals and State of the Art

2.1. Artificial Intelligence and Machine Learning in Surrogate Modeling and Physical Simulations

AI and ML techniques have been increasingly applied to surrogate modeling and physical simulations, resulting in infinstance, the development of physics-informed neural networks (PINNs) demonstrates the potential of ML to solve forward and inverse problems involving nonlinear partial differential equations more efficiently [9]. Furthermore, data-driven approaches have been successfully applied to turbulence modeling, allowing for a better understanding of complex fluid dynamics [10].

2.2. Advances in Surrogate Modeling Techniques

Numerous surrogate modeling techniques have been proposed to represent complex processes and systems, including Gaussian process regression, support vector machines, and deep learning methods [11]. These techniques have been effectively applied to various domains, such as aerospace engineering, materials science, and renewable energy [12]. The growing body of research on surrogate modeling techniques highlights the potential of AI and ML to improve the understanding, optimization, and control of complex processes.

2.3. Developments in Physical Simulations

AI and ML have been increasingly used to advance physical simulations and process optimization in fields such as materials science and manufacturing [13]. For example, researchers have applied ML to optimize additive manufacturing processes, leading to improved product quality and reduced production costs [14]. Additionally, surrogate models have been employed to predict computational fluid dynamics simulations more efficiently, enabling more accurate predictions of complex fluid flow phenomena [15].

The integration of AI and ML techniques into surrogate modeling and physical simulations has the potential to drive further innovation and discovery in various fields.

2.4. Training Data and Model Quality

The quality of surrogate models is highly dependent on the amount and quality of the training data used during the model development process [16]. A sufficient quantity of high-quality data is crucial for ensuring accurate and reliable predictions. However, generating training data for surrogate models, especially for physical simulations, can be computationally expensive and time-consuming [17]. As a result, it is essential to identify the optimal amount of training data necessary for achieving satisfactory model quality without incurring excessive computational costs. Various studies have investigated the relationship between training data volume and model quality. For example, some researchers have examined the effects of using different training data sizes on model performance in the context of fluid dynamics simulations [18]. Additionally, others have explored the impact of data quality on the accuracy of surrogate models in the context of manufacturing processes [19]. These studies demonstrate the importance of understanding the relationship between training data and model quality in order to optimize the development of surrogate models for physical simulations.

3. Methodology

3.1. Simulation of the Production Process

The subsequent analysis of training data for substituting numerical methods with AI-based surrogate models is based on a finite difference analysis of a temperature distribution. Therefore, a more detailed description of the underlying physical model and the boundary conditions is provided. The numerical solution is utilized to generate training data and train the surrogate model. Different AI methods (CNNs, AE, VAE) were employed and evaluated for the surrogate model. In the initial test, the autoencoder [17] showed the most promising results in terms of average error and maximum error. Based on this autoencoder, we developed a feedforward model which showed slightly better results than the autoencoder in further tests, and we ultimately decided to use this architecture. A detailed description of the used architecture, input and output data, and the training process is provided in section 3.3. In the considered scenario, in the glass production, the target is to determine the temperature distribution in the product. Therefore, the underlying PDE in the domain Ω (1) is described through the heat diffusion equation [18].

$$\frac{dT}{dt} = \frac{\lambda}{pc_v}\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}\right) \forall x, y \in \Omega \tag{1}$$

The production scenario represents a continuous process, therefore a transient 2D formulation is used. As for any formulation of a PDE, initial and boundary conditions are necessary to solve the problem. The domain is surrounded by a fluid and adiabatic walls. Hence, on the boundary $\partial\Omega$ (2) of the domain, convection and radiation are the two existing heat transfer phenomena and represent Neumann boundary conditions

$$\frac{dT(x,y,t)}{dx,y} = -\frac{h}{\lambda}(T(x,y,t) - temperature_{air}) - \frac{\varepsilon\sigma}{\lambda}(T(x,y,t)^4 - temperature_{surr}{}^4) \forall x, y \in \partial\Omega \quad (2)$$

Both mechanism are existent on the whole boundary and therefore applied on every edge. At t=0 a homogenous Temperature in the domain is assumed so a initial Temperature is applied. (3)

$$T(x,y,t=0) = T_0 \forall x, y \in \Omega \quad (3)$$

The heat diffusion equation represents an elliptical PDE and therefore, the numerical solution is calculated with finite differences. For this we discretized Ω with constant values Δx,Δy. To obtain the numerical solution we use a forward Euler method as an explicit time integration scheme. Explicit shemes tend to diverge if the time step is too large. Hence, we use the criteria given in [19] for the time stepping criteria. (4)

$$\Delta t \leq \frac{1}{2}\frac{\Delta x^2}{\alpha} \quad (4)$$

In (5) the parameter α describes the diffusion coefficient

$$\alpha = \frac{\lambda}{pc_v} \quad (5)$$

The calculation is performed in a Python 3.11 environment and written in a vectorized form to accelerate the runtime of the numerical solution.

3.2. Training Data Generation

To generate training data for the model, input parameters for the simulation must first be created. To do this, the minimum and maximum value ranges of the parameters that may occur in this production process were determined. The relevant parameters include temperatures, material coefficients, and additional variables such as the product's velocities in the cooling channel. In order to obtain a well-generalized model that covers the various value ranges, the input parameters were randomly generated within the determined value ranges. This ensured that only setting values actually occurring in this production process were used, making the study meaningful and relevant to real-world applications. With these randomly generated input parameters, simulations were then carried out. The results of these simulations, particularly the temperature distribution in the product, were used as training data for the artificial intelligence-based surrogate model. By using a variety of different input parameters and the resulting varying temperature distributions, the model can be applied and evaluated on a wide range of real production conditions. This allows assessing the performance of the surrogate model in terms of accuracy and robustness, ensuring it is suitable for the given application.

3.3. Surrogate Model Design

The surrogate model was developed using PyTorch as the primary framework and is based on a structure partially related to an autoencoder. However, it is ultimately a standalone model. The fully connected neural network, also known as a feedforward network, was designed to model temperature prediction in thermo-mechanical production processes. This surrogate model serves to reduce the computational effort and increase prediction speed for computationally expensive and resource-intensive physical simulations. The model's architecture consists of five linear layers followed by two transposed convolutional layers. The linear layers progressively increase the size of the input data, while the transposed convolutional layers transform the linear output into a spatial temperature distribution. Scaled Exponential Linear Unit (SELU) activation functions are applied after each linear layer. The input for the model includes features relevant to the thermo-mechanical process, such as temperature and time. The model learns from these features and the associated training data to predict the temperature distribution. The output of the final transposed convolutional layer represents the predicted temperature distribution. We chose a feedforward model with convolutional layers for the surrogate model, as it outperformed other model types, such as Convolutional Neural Networks (CNNs), Autoencoders (AE), and Variational Autoencoders (VAE), in our tests. The rationale for this choice is based on several factors: i. Simplicity and efficiency: Feedforward models have a simpler structure and require less computational power compared to CNNs, AEs, and VAEs. In this specific use case, a simpler model is sufficient to produce meaningful results when analyzing the relationship between the training data volume and model quality. ii. Scalability: Feedforward models are easily scalable and can be adapted to various data volumes and complexity levels. This allows for better exploration of the relationship between training data volume and model quality, as the model can flexibly respond to different data sizes. iii. Relevance for physical simulations: In physical simulations, the underlying processes are often time-invariant and linear or nearly linear. Feedforward models are well-suited to represent such relationships, as they can capture linear and non-linear relationships without unnecessary model complexity. Overall, we decided on the feedforward model mainly due to the best performance. However, simplicity, efficiency, scalability, and relevance for physical simulations also played a crucial role in our decision.

3.4. Model Training Approach

Before training the surrogate model, we performed a brief hyperparameter tuning to determine the optimal values for several parameters: batch size, learning rate, maximum epochs, and train-test split rate. These values were then fixed for subsequent analyses to ensure consistent results during the evaluation process. Following the hyperparameter tuning, we trained the model with varying amounts of training data to investigate its performance under different data set sizes. We started with 50 training samples and gradually increased the number in increments of 50, stopping at 1,000 samples. This approach allowed us to explore the impact of the training data size on the model's prediction accuracy and generalization capabilities. To ensure the robustness of our findings, we repeated the training process multiple times with different training data quantities. This helped in understanding how the model's performance would vary with the size of the training data and provided valuable insights into the optimal training data size for the given application.

4. Evaluation and Results

4.1. Model Quality Metrics

In this chapter, we discuss the evaluation results and performance indicators obtained for the trained models. We determined both the Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE) as performance metrics for all trained models at different step sizes and training runs. Two plots were created to visualize the quality of the machine learning model: Fig 1. (a) shows the adjusted MAE, and Fig 1. (b) displays the adjusted RSMAE as a function of the training data. We used MAE and RMSE metrics for the following reasons: They allow us to evaluate the quality of the model by calculating the deviations between the predicted and actual values. The lower the MAE and RMSE, the better the model's ability to generalize the data and make accurate predictions. In our study, we chose the train-test-split method as the validation technique, as it is best suited for our use case. We opted for this method for the following reasons: In our study, we use a limited dataset of 1,000 training examples. However, cross-validation and bootstrapping require the dataset to be split multiple times, leading to a further reduction in data volume. This may result in the model not being adequately trained, potentially leading to poorer model quality. The train-test-split method divides the dataset only once into a training and testing dataset, making the entire data volume available for model training. This allows the model to be better trained, potentially leading to improved model quality. The train-test-split method is widely used in practice and well-understood. It is also easier to implement and comprehend than cross-validation or bootstrapping. Overall, the train-test-split method is a suitable technique for our use case, as it provides an adequate amount of training data and is easy to implement, leading to an effective investigation of the relationship between training data volume and model quality.

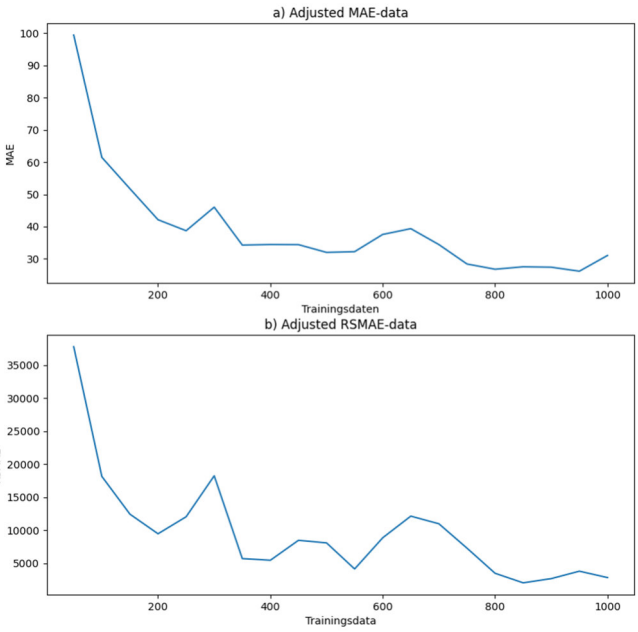

Fig. 1. (a) Adjusted MAE-data; (b) Adjusted RSMAE-data.

4.2. Influence of Training Data Quantity on Model Quality

The quality and accuracy of AI models depend heavily on the amount and quality of the training data used to train the models. In many AI applications, there are large amounts of training data available in accessible databases and can be easily extended. However, generating training data for surrogate models of physics simulations is computationally demanding and requires a significant amount of time and resources. To determine the necessary amount of training data required to build a good AI model for a thermomechanical production process, a study was conducted with different sizes of artificially generated training data. The study involved testing the performance of a surrogate model using different amounts of training data, ranging from 50 to 1000. The results of the study showed that the mean absolute error (MAE) decreased as the amount of training data increased up to a certain point, after which it started to increase again. The lowest MAE value of 26.14 was achieved with 950 training data points. However, the root mean squared absolute error (RSMAE) showed a different trend. It continued to decrease as the amount of training data increased, suggesting that the model's predictive power improved with an increase in training data. In summary, the study confirms that the amount of training data significantly influences the quality and accuracy of AI models. It also shows that there is a threshold for the amount of training data needed to achieve good model performance. While increasing the amount of training data can improve the model's predictive power, there is a point at which the returns diminish. Therefore, careful consideration should be given to the amount and quality of the training data used in building AI models for complex systems, such as physics simulations, to ensure optimal model performance.

5. Conclusion and Future Work

The application of artificial intelligence (AI) and machine learning (ML) has rapidly advanced in various fields in recent years. Particularly notable are the developments in the modeling and simulation of physical processes. Surrogate models, also known as metamodels, have the potential to capture complex relationships between input and output parameters and significantly reduce computation time compared to traditional simulations. This also provides the possibility to use such methods for evaluation and recommendation in production environments that were previously not possible due to their complexity or computation time. However, the quality of these models depends on the quantity and quality of available training data. Generating training data for surrogate models based on physical process models is expensive and computationally intensive. Therefore, it is advantageous to know the necessary amount of training data beforehand. In this work, a flat glass production process was simulated to gain an understanding of the process and make predictions about the temperature profile during the production process in the product. The goal was to replicate this in a surrogate model that can make predictions much faster and work with a relatively short response time. This offers the possibility to create a control system for complex production processes. For this purpose, training data sets were generated, and in a next step, a suitable model architecture was identified. Finally, the selected model was trained with different amounts of training data to establish a relationship between the amount of training data and the model quality. A detailed presentation of the methods and techniques used was presented in Sections 2 and 3. In Section 4, the results of the study were presented, with a particular focus on analyzing the effects of the amount of training data on the model quality. The results showed that the mean absolute error (MAE) decreased as the amount of training data increased up to a certain point, but then began to increase again. The lowest MAE value of 26.14 was achieved with 950 training data points. However, the Root Mean Squared Absolute Error (RSMAE) showed a different trend. It continued to decrease as the amount of training data increased, indicating that the predictive power of the model improved with an increasing number of training data. Overall, the study showed that the quantity and quality of training data have a significant impact on the quality and accuracy of AI models. It was also shown that there is a threshold for the required amount of training data to achieve good model performance. While increasing the amount of training data can improve the predictive power of the model, there is a point of diminishing returns. Therefore, when creating AI models for complex systems such as physical simulations, careful consideration should be given to the quantity and quality of training data to ensure optimal model performance. Future research could focus on investigating other AI methods for creating surrogate models, such as decision trees or random forests. Another possibility is to extend the study to other thermomechanical production processes to test the generalizability of the results. It would also be interesting to further investigate the relationship between the quality of training data and model performance to determine how improving the quality of training data affects model performance. Finally, the integration of online learning methods into surrogate models for more efficient optimization of production processes could be explore

References

[1] Fujii, H., Teshima, Y., Funabiki, S., & Hoshino, T. (2020). Machine learning for fluid dynamics: A review of data-driven turbulence modeling. Physics of Fluids, 32(12), 121301.

[2] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707.

[3] Jin, H., Yang, G., Wang, G., & Karniadakis, G. E. (2021). Surrogate modeling for fluid flows based on physics-informed deep learning: A review. Applied Mechanics Reviews, 73(1), 010801.

[4] Lüthen, N., Tewolde, M. H., Führer, C., & Körner, C. (2021). Simulation-based optimization of laser powder bed fusion processes using artificial neural networks. Materials & Design, 204, 109724.

[5] Wang, Y., Sun, J., & Kudo, M. (2021). Data-driven methods for modeling and optimization of manufacturing processes: A review. International Journal of Machine Tools and Manufacture, 169, 103724.

[6] Yang, Y., Zhang, R., & Halem, M. (2020). Deep learning based surrogate models for prediction of computational fluid dynamics simulations. Computers & Fluids, 200, 104447.

[7] Beck, J., & Stengel, R. F. (2019). Machine learning for aircraft trajectory prediction in cruise. Journal of Guidance, Control, and Dynamics, 42(5), 1169-1179.

[8] Benham, R., & Hjorth, L. (2021). Machine learning for molecular dynamics in chemical engineering. Chemical Engineering Science, 236, 116468.

[9] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686-707.

[10] Ling, J., Kurzawski, A. K., & Templeton, J. (2016). Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. Journal of Fluid Mechanics, 807, 155-166.

[11] Forrester, A. I., & Keane, A. J. (2009). Recent advances in surrogate-based optimization. Progress in Aerospace Sciences, 45(1-3), 50-79.

[12] Palar, P. S., Tao, Z., & Rahman, S. (2020). A review of machine learning applications in renewable energy. Renewable Energy, 145, 1119-1132.

[13] Vlachogiannis, D., & Siettos, C. (2018). Machine learning accelerated computational fluid dynamics. Computers & Fluids, 175, 84-92.

[14] Li, H., & Li, Y. (2020). A review on machine learning in additive manufacturing: Data collection, methods, and applications. Materials & Design, 195, 108966.

[15] Yang, Y., Zhang, R., & Halem, M. (2020). Deep learning based surrogate models for prediction of computational fluid dynamics simulations. Computers & Fluids, 200, 104447.

[16] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345-1359.

[17] Marzouk, Y. M., & Willcox, K. E. (2014). Exploiting active subspaces to quantify uncertainty in the numerical simulation of the HyShot II scramjet. Journal of Computational Physics, 302, 1-20.

[18] Renganathan, V., & Ilangovan, R. (2017). Study on the effect of training dataset size on prediction accuracy of artificial neural network surrogate models for CFD simulations. Procedia Engineering, 181, 713-720.

[19] Hämäläinen, J., Špakov, O., & Rauhala, U. (2015). Effects of sample size and dimensionality on the quality of data-based aping models. Expert Systems with Applications, 42(21), 7413-7423.

[20] Kim, D-J., Kim, S-I., & Kim, H-S. (2022). Thermal simulation trained deep neural networks for fast and accurate prediction of thermal distribution and heat losses of building structures. Applied Thermal Engineering, 202, 117908. ISSN 1359-4311.

[21] Baehr, H. D., & Stephan, K. (2019). Wärme- und Stoffübertragung. Springer Vieweg Berlin, Heidelberg. doi:10.1007/978-3-662-58441-5.
[22] Pearson Studium. (2009). Maschinenelemente: Band 1: Konstruktion und Berechnung von Verbindungen, Lagern, Wellen (2nd updated edition) [Publisher number:: 326670, eBook (PDF), ISBN: 978-3-86326-670-7]. Pearson Studium - Mechanical Engineering.
[23] Zhang, W., Choromanska, A., & LeCun, Y. (2020). Deep learning. Natur