Journal of Computer Science
https://dergipark.org.tr/en/pub/bbd

ISSN,e-ISSN: 2548-1304
Volume: 2026
https://doi.org/**10.53070/bbd.1992059**
Research Paper

Machine Learning-Based Surrogate Modeling and Genetic Algorithm Optimization of Industrial Spray Drying Operations

Endüstriyel Sprey Kurutma Operasyonlarının Makine Öğrenmesi Tabanlı Vekil Modellemesi ve Genetik Algoritma ile Optimizasyonu

Lawrence A. Farinola[*1]

[1]Department of Software Engineering, Faculty of Architecture and Engineering, Rauf Denktas University, Mersin 10 via Turkey.

(lawrence.farinola@rdu.edu.tr)

Received: Jul.11, 2026 Accepted: Aug. 18, 2026 Published: Aug.18, 2026

Özetçe— Endüstriyel sprey kurutma, termal, mekanik, çevresel ve zamansal değişkenler arasındaki etkileşimlerle karakterize edilen karmaşık doğrusal olmayan bir üretim sürecidir. Bu durum, süreç performansının doğru biçimde tahmin edilmesini ve çalışma koşullarının optimize edilmesini zorlaştırmaktadır. Bu çalışmada, endüstriyel sprey kurutma sürecinin iyileştirilmiş çalışma koşullarının belirlenmesi amacıyla Rastgele Orman Regresyonunu (RFR) vekil süreç modeli olarak kullanan ve Genetik Algoritma (GA) ile birleştirilen veri odaklı bir optimizasyon çerçevesi geliştirilmiştir. RFR modeli, süreç değişkenleri ile seçilen süreç performans göstergesi arasındaki doğrusal olmayan ilişkileri modellemek amacıyla geliştirilirken, GA tanımlanan süreç kısıtları altında uygulanabilir çalışma uzayında daha uygun operasyonel konfigürasyonları araştırmak için kullanılmıştır. Tahmin modeli, mevcut endüstriyel gözlemler üzerinde güçlü bir tahmin performansı göstermiştir. Özellik önem analizi, gaz akışı ve zamansal değişkenlerin süreç performansıyla ilişkili etkili değişkenler arasında olduğunu göstermiştir. GA tabanlı optimizasyon, tanımlanan operasyonel kısıtları koruyarak temel çalışma koşuluna kıyasla model tarafından tanımlanan süreç hedefinde yaklaşık %5,91 oranında iyileşme ile ilişkili bir çalışma konfigürasyonu belirlemiştir. Bulgular, veri odaklı vekil modellemenin evrimsel arama ile bütünleştirilmesinin endüstriyel sprey kurutma süreçlerinin optimizasyonunda pratik bir potansiyele sahip olduğunu göstermektedir. Çalışmanın katkısı, RFR ve GA kombinasyonunu yeni bir optimizasyon algoritması olarak sunmaktan ziyade, tahmine dayalı süreç modellemesini, yorumlanabilir değişken analizini ve kısıtlı operasyonel optimizasyonu Endüstri 4.0 üretim ortamında bütünleştiren uygulama odaklı bir çerçeve geliştirilmesidir.

Anahtar Kelimeler : *Makine Öğrenmesi; Rastgele Orman Regresyonu; Genetik Algoritma; Vekil Modellemesi; Sprey Kurutma; Süreç Optimizasyonu; Endüstri 4.0.*

Abstract— Industrial spray drying is a complex nonlinear manufacturing process characterized by interactions among thermal, mechanical, environmental, and temporal variables, making accurate process prediction and operational optimization challenging. This study develops a data-driven optimization framework that uses Random Forest Regression (RFR) as a surrogate process model and a Genetic Algorithm (GA) to identify improved operating configurations for industrial spray drying. The RFR model was developed to capture nonlinear relationships between process variables and the selected process-performance response, while the GA was used to search the feasible operating space subject to the defined process constraints. The predictive model demonstrated strong predictive performance across the available industrial observations. Feature-importance analysis identified gas-flow and temporal variables among the influential predictors, providing additional insight into the factors associated with process performance. The GA-based optimization identified an operating configuration associated with an approximately 5.91% improvement in the model-defined process objective relative to the baseline condition while satisfying the specified operational constraints. These findings demonstrate the practical potential of combining data-driven surrogate modeling with evolutionary search for process optimization in industrial spray drying. Rather than treating the combination of RFR and GA as a novel optimization algorithm, the study contributes an application-oriented framework for integrating predictive process modeling, interpretable variable analysis, and constrained operational optimization within an Industry 4.0 manufacturing environment.

Keywords : *Machine Learning; Random Forest Regression; Genetic Algorithm; Surrogate Modeling; Spray Drying; Process Optimization; Industry 4.0.*

1. Introduction

With the swift advancement of Industry 4.0, there has been an increase in the incorporation of technology such as machine learning (ML), cyber-physical systems, industrial sensing, and decision support systems based on data. All these technological advances are important since manufacturing can move from simple monitoring to predictive analysis, process control, and optimization of processes through the use of these systems. ML is highly essential in the current scenario where industrial processes have a lot of nonlinearity, changing process conditions, and interaction between various process variables that cannot be modeled using traditional approaches (Jordan & Mitchell, 2015; Wuest et al., 2019). Modern advances in smart manufacturing have underscored the importance of integrating predictive modeling and operational decision support systems (Farinola, 2026).

An example of such a highly complex manufacturing setting is industrial spray drying. Spray drying is utilized in food, pharmaceutical, chemical, and related industries, although the process is characterized by significant interdependencies between thermal, mechanical, environmental, and materials-related variables. Such process parameters as drying air flow, temperature, humidity, pressure, and process duration can affect process efficiency and product properties at the same time. Thus, process modifications will have a cascading effect on other process variables in a nonlinear fashion, making it difficult to predict and optimize the process. Since spray drying is an energy-consuming process, it becomes even more important to detect process settings that allow enhancing efficiency without sacrificing process feasibility. In recent reviews, machine learning has been highlighted as one of the key modeling tools for spray drying due to its ability to model complex interdependencies between process variables without a full first-principles understanding of the system (Wahab et al., 2025).

Traditional optimization of industrial drying processes involves the use of either empirical tests, experience-based optimization, response surface modeling, or iterative physical tests. While such approaches may be effective within a well-controlled environment, their effectiveness may be limited in cases where there are many interacting parameters to take into account at once. This is where machine learning can prove useful by finding patterns and relationships between input variables and performance parameters using historic operational data. The Random Forest Regression (RFR) algorithm is especially useful since it can find non-linear relationships between input variables, consider interactions among the predictors, measure the importance of variables, and make relatively limited assumptions regarding the functional form (Breiman, 2001). Previous literature has already found machine learning to be applicable in modeling spray-drying and other drying processes.

Nevertheless, prediction is not enough for obtaining an operational decision. It is possible for a model to make accurate predictions about process performance but still fail to determine what combination of controllable variables is required in order to optimize the performance. The difference between predictive modeling and prescriptive process optimization becomes significant at this point. Evolutionary algorithms, such as Genetic Algorithm (GA), can solve this issue because they are able to perform search in the large and nonlinear space of decisions in accordance with pre-specified objective functions and constraints (Deb et al., 2002). The model, in the case of surrogate-assisted optimization, can act as a relatively cheap representation of the actual process and, thus, a lot of different operating conditions can be considered without performing any new experiments.

It is important to note, however, that the use of machine learning combined with a Genetic Algorithm is not per se an innovative approach. The previous literature includes cases where machine learning models were combined with evolutionary optimization techniques in order to optimize spray drying operations. In particular, ANN and GA have been combined in modeling and optimizing the spray drying process, showing that it is possible to combine learning with evolution. Even more recently, researchers have studied surrogate-based spray dryer optimization using Random Forest and neural networks in combination with GA, Particle Swarm Optimization, and evolutionary hybrids. It follows, therefore, that the novelty of the current research does not lie in the proposed RF-GA combination technique. Rather, the research aims at implementing and assessing a decision-making framework based on surrogates in actual industry spray drying time series data.

Another issue related to the selection of the prediction and optimization algorithms. In modern industrial machine learning research, competitive ensemble techniques such as Random Forest, XGBoost, LightGBM, and CatBoost are often used for prediction; meanwhile, for optimization problems, techniques like GA, Particle Swarm Optimization (PSO), simulated annealing, and other metaheuristics are often taken into account. Therefore, the selection of RFR and GA for the present research should be considered as a methodological decision and be justified rather than the assumption of their overall superiority. RFR was chosen as the main surrogate because of the presence of nonlinear relationships between process variables in the dataset and also due to the trade-off of predictive ability, robustness, efficiency, and explainability through the feature importance. GA was chosen because the optimization problem deals with bounded operational variables and potentially nonlinear objective functions; hence, the use of evolutionary search is quite appropriate for searching the operational space.

Recent research findings also prove that more sophisticated methods can provide promising outcomes in the spray drying optimization process. Digital twin-enabled transfer learning has been explored to enhance energy efficiency monitoring for scenarios when real process data is limited, whereas recent multi-objective spray drying research has compared predictive models and optimization algorithms. All of these advancements suggest that the modern research agenda is gradually moving away from proving that ML can predict the behavior of the spray drying process towards exploring how such predictions can be translated into decisions.

In this context, there is an obvious need for the incorporation of real industrial operating data, predictive surrogate modelling, process interpretability, and constrained optimization into a unified decision support approach. Current research has often dealt with experimental data from laboratory or pilot scale, simulation data, formula-specific optimizations, or predictive tasks alone. While other research exhibits strong predictive capabilities, they lack the aspect of translating their predictions into operations. On the other hand, while optimization techniques find the optimal solution in a mathematical sense, they fail to validate the predictive abilities of the surrogate model employed in evaluating the candidate solutions. Current reviews highlight the dependence of machine learning applications in spray drying on good quality data, process variability, generalization of models, and linking predictive models to process control and optimization.

This study aims to fill this gap through 1,503 operational data from industrial spray-drying processes and the creation of surrogate-based approach where Random Forest Regression models behaviour of process and Genetic Algorithm finds better operating regimes in the given ranges. In this study, the predictive ability of the surrogate model is checked, the importance of the process variables is studied, and the created model is used as a surrogate to represent the process for evolutionary optimization. Therefore, the focus of this work lies in the complete chain of analysis: from industrial data to predictive modeling to model interpretation to constrained optimization to operational recommendation.

The main contributions of the paper include:

1. Process data-driven surrogate modeling: In this paper, a Random Forest Regression surrogate is developed and evaluated with the use of 1,503 real process data points for representing the relationship between process variables and performance.
2. From predictions to prescriptions: In this paper, it is shown that a predictive surrogate can be linked with a Genetic Algorithm for turning process predictions into viable operating conditions.
3. Interpretation of optimization results: It was considered feature importance analysis as a part of interpretability for understanding which process variables play an important role in process performance and evaluation of optimization suggestions.
4. Evaluation of an industrial optimization workflow: The proposed approach is evaluated based on historical industrial process data and constraints on operating conditions..

Accordingly, the objective of this study is not to claim that Random Forest Regression combined with a Genetic Algorithm constitutes a fundamentally new optimization algorithm, but to investigate whether this surrogate-assisted workflow can provide accurate, interpretable, and operationally useful decision support for a nonlinear industrial spray-drying process under real-world data conditions.

Research Questions

RQ1: How accurately can Random Forest Regression model the nonlinear relationship between operational variables and process performance in an industrial spray-drying system?

RQ2: Can a Random Forest surrogate integrated with a Genetic Algorithm identify operational configurations that improve process performance relative to the observed baseline conditions while satisfying predefined operational constraints?

RQ3: Which process variables contribute most strongly to the predictive behaviour of the surrogate model, and how can this information support interpretation of the resulting operational recommendations?

RQ4: To what extent does the proposed surrogate-assisted workflow provide a practically interpretable decision-support mechanism for industrial spray-drying optimization?

2. Literature Review

2.1 Machine Learning in Industrial Manufacturing

Industry 4.0 is resulting in increased deployment of machine learning (ML), cyber-physical systems, industrial sensors, and data analytics in the manufacturing sector. The aforementioned technologies help with predictive maintenance, quality monitoring, process prediction, anomaly detection, and process optimization through identification of relationships using industrial data (Jordan & Mitchell, 2015; Wuest et al., 2019; Lee et al., 2015). As compared to traditional linear or rule-based models, machine learning algorithms have the ability to capture non-linear relations and interactions between different process variables in the manufacturing setting.

There have been many machine learning techniques used for industrial prediction, such as ensemble learning, support-vector-based techniques, and deep learning. However, the choice of the model should be based on the nature of the data at hand, the goal of prediction, computation capabilities, and interpretability

needs. Random Forest Regression (RFR) is especially appealing in the field of industry since it allows for modeling nonlinear relations, interaction of variables and provides a measure of variable importance without the need for a pre-defined function form (Breiman, 2001). Nonetheless, RFR is not always better than other ensemble techniques like XGBoost, LightGBM, or CatBoost. The latter can outperform RFR on some kinds of data structures. Thus, the choice of the RFR model in this research is justified by its ability to predict nonlinearly, robustness, relative simplicity, and applicability in conjunction with process interpretation based on feature importance measures.

Predictive power of a model per se, however, does not always give the basis for industrial decisions. In cases when operation variables should be chosen within certain intervals and the goals of the process should be optimized at the same time, the connection of the prediction model with the optimization algorithm becomes inevitable.

2.2 Predictive Modeling in Spray Drying Systems

As a manufacturing process that is thermally intensive and involves complex interactions between various thermal, environmental, mechanical, and temporal variables, spray drying can be optimized in relation to process efficiency and performance by examining nonlinear relationships (Farinola & Bazarkhan, 2025).

Machine learning offers an approach to the problem, in which process relations are learnt based on historical operational data. Among several approaches, RFR stands out for its ability to deal with both nonlinear effects and interaction of different variables while being quite insensitive to industrial measurement noise (Breiman, 2001). Another benefit offered by the method is its feature-importance estimation, which is especially useful in case the purpose is not only to forecast process performance but also to understand what variables impact this performance the most.

Industrial applications need interpretability of optimization recommendations in order to ensure their operability and defendability. Basic feature-importance analysis can serve as a first-level interpretation of variables involved in model prediction, while other advanced methods such as SHAP offer local and global explanations of the predictions (Lundberg & Lee, 2017). In this study, feature-importance analysis is used to complement the predictive models to identify the dominant process variables.

2.3 Evolutionary Optimization in Manufacturing Systems

While the predictive model predicts the behavior of the process, it does not automatically dictate what the best operating condition should be. Evolutionary optimization algorithms offer a way of exploring complicated decision spaces, where the objective function is nonlinear, discontinuous, or not easy to optimize. Genetic Algorithm (GA), which makes use of an evolutionary process in a population, has been extensively used in engineering and manufacturing optimization problems (Holland, 1975; Deb et al., 2002).

GA is selected in this study because the spray-drying optimization problem involves multiple bounded operational variables and a nonlinear surrogate objective. The population-based search mechanism allows candidate operating configurations to be evaluated without requiring gradient information from the underlying industrial process. This is particularly useful when the optimization is performed through a machine-learning surrogate rather than through a differentiable physical process model.

Nevertheless, GA is only one of several suitable metaheuristic approaches. Methods such as Particle Swarm Optimization (PSO), Differential Evolution, Simulated Annealing, and other evolutionary algorithms may also be appropriate for nonlinear industrial optimization. Therefore, the use of GA in this study should not be interpreted as a claim that it is universally superior to other optimization algorithms. Rather, GA provides a practical and established search mechanism for evaluating the feasibility of surrogate-assisted operational optimization in the specific industrial dataset examined here.

2.4 Machine Learning–Optimization and Surrogate-Assisted Systems

The integration of predictive models with optimization algorithms has become an important approach for data-driven industrial decision support. In surrogate-assisted optimization, a computational model approximates the behaviour of a physical or otherwise expensive process, allowing large numbers of candidate operating configurations to be evaluated at substantially lower computational cost. This approach is particularly useful when direct experimentation is expensive, time-consuming, or operationally disruptive.

Previous research has demonstrated the feasibility of combining machine learning models with evolutionary optimization for drying and spray-drying applications. Such studies indicate that predictive models can be used not only to estimate process outcomes but also to search for operating conditions associated with improved performance. However, the existence of ML–GA or other ML–optimization combinations means that the combination of RFR and GA alone should not be regarded as the principal methodological novelty of the present study.

Instead, the relevance of the present framework lies in connecting four components within a single workflow: real industrial process data, predictive surrogate modelling, interpretable process analysis, and

constrained evolutionary search. This structure enables the predictive model to function as a computational representation of the industrial process while the optimization algorithm searches the feasible operating space. Feature-importance analysis provides an additional layer through which the resulting recommendations can be interpreted in relation to the underlying process variables.

This distinction is important because industrial deployment requires more than a high prediction score. An operational framework must also provide feasible recommendations, identify the variables influencing the predicted outcome, and allow the relationship between historical process behaviour and optimized conditions to be examined.

2.5 Research Gap

Existing research shows the potential of machine learning for predictive modeling of industrial processes and evolution strategies for optimization. There are, however, some limitations that affect industrial spray-drying decision support in particular.

Firstly, prediction and optimization problems are often assessed separately as two different analysis problems. While one problem involves obtaining accurate predictions from a model, another problem involves finding solutions to optimization equations and ignoring the accuracy of predictions obtained by the model.

Secondly, interpretation of the results remains significant, although it is not always incorporated in the combined approach. Users must be aware of the factors that affect model performance before making any algorithm-based operating suggestions.

Finally, despite the fact that combinations of machine learning and optimization techniques have already been studied in the case of drying and similar engineering problems, further empirical studies based on actual industrial data are still required.

Ultimately, the choice of the methodology itself needs to be empirically justified. RFR is not better than XGBoost, LightGBM, or CatBoost, and GA is not better than PSO or any other optimization algorithm based on the idea of metaheuristics. Thus, the value of this research is not the creation of a novel approach to ML–optimization, but rather a development and validation of a surrogate-based industrial decision-support process that uses RFR for nonlinear process prediction and interpretation, while GA explores the feasible design space.

This study tests if there is a viable route from industry process data to predictions, interpretations, and optimization within such an approach. The test is performed using 1,503 real industrial spray-drying data samples, focusing on the quality of the prediction, feature importance, operability, and the degree of improvement compared to the baseline operating conditions.

3. Methodology

3.1 Research Design and Computational Framework

This research uses a computationally-driven research methodology to explore the use of Random Forest Regression (RFR) as a surrogate modeling technique for the optimization of an industrial spray drying process. The research methodology comprises four stages, which include: (i) preparation and pre-processing of the industrial data, (ii) construction and validation of the RFR prediction model, (iii) feature importance analysis for the interpretation of the model, and (iv) optimization of the process using the Genetic Algorithm (GA).

The key research methodology involved the use of the trained RFR model as the surrogate model of the observed industrial process. The surrogate model predicts the process responses for different operating conditions, while the GA algorithm searches for optimal configurations of the process within the feasible operational region. The GA search process avoids conducting multiple experiments on the real process during the computation-based optimization process.

Predictive performance was analyzed using regression measures and time-based validation in order to investigate the stability of the proposed surrogate model. Feature Importance Analysis was later applied to determine which process variables have the highest influence on the predictive behavior of the model. The obtained surrogate model was then used in the GA in order to find an optimal operating configuration under certain operational constraints. The obtained configuration was compared to the baseline operating point based on the optimization objective and the percentage of improvement.

It is important to note that neither Random Forest Regression nor Genetic Algorithms are presented in this paper as novel algorithms. They are well-known techniques. Methodological novelty, rather, is provided by the use of the mentioned algorithms in a structured approach of surrogate model construction, analysis and optimization based on industrial observations of spray-drying process.

3.2 Industrial Case Study and Data Description

The methodology was evaluated using operational data obtained from an industrial spray-drying process. Spray drying represents a nonlinear manufacturing environment in which thermal, environmental, mechanical, and temporal variables interact and influence process behaviour.

The original dataset consisted of 1,510 time-series observations collected over a six-month operating period. Following data-quality screening and removal of incomplete records, 1,503 valid observations were retained for the analysis. The resulting dataset represents the operational conditions observed during the study period rather than experimentally generated observations.

The predictor variables comprised operational, environmental, and temporal characteristics of the spray-drying process, including gas flow rate, inlet temperature, humidity, pressure, outside temperature, and temporal process indicators. The process response used for predictive modelling was the outlet temperature, which was treated as the selected process-performance response for the optimization analysis.

The use of historical industrial observations provides an empirical basis for evaluating whether the proposed predictive–optimization workflow can identify potentially improved operating configurations under the range of conditions represented in the available process data.

3.3 Data Preprocessing and Feature Engineering

A structured preprocessing procedure was applied before model development. Records with missing values in the target variable were removed, and the completeness of the predictor variables was examined. Data-quality screening also considered anomalous observations using interquartile-range criteria and process-specific thresholds.

Temporal information was retained rather than randomly shuffled because the observations represent sequential industrial operations. Preserving the temporal structure was important for reducing the risk of information leakage between training and testing observations.

The final dataset contained 1,503 valid observations. The predictors were subsequently prepared for model development and optimization. Although Random Forest Regression does not require feature scaling for its tree-based splitting mechanism, normalization was applied to the operational variables used within the optimization procedure to provide a consistent numerical representation of the search variables.

Temporal indicators were retained as candidate predictors because process behaviour may vary according to operational time and changing environmental conditions. No target-derived variables were introduced into the predictor set.

3.4 Random Forest Regression Model

Random Forest Regression was selected as the surrogate modelling technique because of its ability to represent nonlinear relationships and interactions among process variables without requiring a predefined functional form. The method is also suitable for heterogeneous industrial observations and provides a direct mechanism for assessing predictor importance through the trained ensemble.

A Random Forest regression model consists of an ensemble of decision trees whose individual predictions are aggregated to obtain the final prediction:

$$\hat{y} = \frac{1}{N}\sum_{i=1}^{N} f_i(x)$$

where $f_i(x)$ represents the prediction generated by the i^{th} regression tree and N denotes the number of trees in the ensemble.

The model was trained with the help of 100 estimators along with restricted tree-depth and minimum-split settings. Hyperparameter tuning included the number of estimators, maximum tree depth, and minimum samples necessary for a split. Model selection was carried out according to the performance of the model in terms of predictive capabilities under time-aware validation.

RFR was chosen as the main surrogate rather than claiming its superiority to other ensembles. Algorithms like XGBoost, LightGBM, and CatBoost are comparable alternatives, but this current research concentrates on whether RFR will be an adequate and sufficiently fast surrogate in the context of the particular industrial dataset at hand. The reason behind choosing this algorithm is the fact that it can handle non-linearities, is noise-resistant, has moderate computational needs, and can be analyzed via feature importance for the purpose of interpretation.

Thus, the exclusion of other machine learning techniques from this current investigation should not be taken as the indicator of RFR's superiority to other algorithms.

3.5 Training, Testing, and Time-Aware Validation

Since the data was time-series data, a temporal partitioning of the data was made, as opposed to random split between training and test sets. The first 80% of the observations were used for training the models, whereas the

last 20% of the observations were held out for testing. This approach keeps the order of the data, and gives a more realistic estimate of the performance of predictions based on future observations.

Apart from temporal validation of the data, time-series cross validation technique was adopted to investigate the stability of model performance in various time periods. It was done to ensure that the performance of the test set was not a function of a specific portion of the data set.

Performance of the model was measured in terms of coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE). These metrics provided complementary insights into performance of prediction and explanation of variance.

Coefficient of determination is computed as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean observed value.

Mean absolute error was calculated as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i - \hat{y}_i|$$

while root mean squared error was calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The combination of chronological testing and time-aware cross-validation was used to provide evidence regarding the predictive reliability of the surrogate before it was employed in the optimization stage.

3.6 Feature-importance Analysis

A feature importance analysis was included in the investigation to account for the interpretability element in the study. Feature importance values of the predictor variables were calculated following the fitting of the Random Forest model.

The approach allowed us to detect which variables had the highest importance in the behavior of predictions made by the RFR model. This kind of information serves as an interpretable way to verify if the variables that are important for prediction are also relevant to the operational configurations determined by the optimization method.

Feature importance was considered to be an interpretative element and not a causal one. In other words, high importance implies that a particular variable plays a significant role in the prediction structure generated by the Random Forest model; it does not imply that the change in such a variable leads to changes in process responses.

The reason why it is important to distinguish between these two concepts is that in the context of industry, our main concern is to interpret model-generated conclusions and not to deduce any causal relations from the process data observations alone.

3.7 Genetic Algorithm Optimization

The second computational component of the framework was a Genetic Algorithm used to search for improved operating configurations within the feasible process space.

Each candidate solution was represented as a chromosome containing the selected operational variables:

$$C = (x_1, x_2, \dots, x_n),$$

where x_j represents an operational decision variable.

The trained Random Forest model was used as the surrogate fitness evaluator. The optimization problem can therefore be expressed generally as:

$$\max_{x \in \Omega} F(x) = f_{RFR}(x), \qquad \text{subject to } l_j \le x_j \le u_j$$

where $f_{RFR}(x)$ represents the process response predicted by the Random Forest model, Ω denotes the feasible operating region, and l_j and u_j represent the lower and upper bounds for each operational variable.

GA was set up with a population size of 50 candidate solutions and run for up to 120 generations. The crossover rate and mutation rate were set at 0.80 and 0.10 respectively. Tournament selection was utilized for the selection of parents, whereas elitism was used as part of the algorithm to retain good solutions from generation to generation.

Operating limits were incorporated within the optimization process to ensure that the algorithm does not create any solution beyond the operable space. The optimization process stopped either upon reaching the

maximum number of iterations or when there is no improvement in the fitness of the best solution within the pre-defined convergence criterion for the specified number of consecutive generations.

GA was chosen as the optimization tool because the current problem deals with bounded operational variables and a nonlinear surrogate function that does not need gradient information. Alternative optimization algorithms, such as Particle Swarm Optimization, Differential Evolution, and Simulated Annealing, can also be used on this type of problem. Nevertheless, the comparison of several different optimizers was out of the scope of the current study.

3.8 Baseline and Optimization Evaluation

The optimization process was performed by comparing the performance measure linked with the baseline operating condition against the performance measure of the optimized solution generated via GA.

The baseline operating condition is the reference operating condition against which the computationally optimized operating condition has been compared. The computationally optimized operating condition has been derived with respect to the constraints and then has been tested using the trained surrogate model based on Random Forest. The percentage improvement is calculated as:

$$Improvement\ (\%) = \frac{F_{opt} - F_{base}}{F_{base}} \times 100,$$

where F_{opt} denotes the optimized process-performance value and F_{base} denotes the corresponding baseline value. The baseline-to-optimized comparison was used to determine whether the computational optimization procedure identified a configuration associated with improved process performance relative to the observed reference condition. Because the study evaluates a single optimization trajectory and does not have an independently generated set of repeated baseline–optimized observations, the improvement is interpreted descriptively rather than as an inferential estimate of population-level statistical significance.

3.9 Integration of the Predictive and Optimization Components

The entire computational pipeline was based on an analytical path from industrial observations to operational recommendations. Initially, the industrial spray-drying observations were collected and filtered to ensure data quality, yielding a total number of 1,503 valid records. The time structure of the observations was retained in the modeling phase, and the data were split chronologically into the training and test datasets.

In the next step, a Random Forest Regression model was trained based on the selected operational variables and tested using R^2, MAE, RMSE, and time-sensitive cross-validation. Then, feature importance was analyzed to study the role of the predictor variables in the constructed model.

In the next step, the trained Random Forest model was utilized as a fitness function of the Genetic Algorithm. The Genetic Algorithm explored the pre-defined feasible region of operational variables and created candidate solutions according to the objective function and constraints. The most efficient solution was compared against the baseline operating point to estimate the possible improvement in the process performance.

This procedure creates a direct analytical path from industrial data → predictive modeling → model interpretation → constrained optimization → operational recommendation, while maintaining a distinction between the steps of prediction, interpretation, and optimization.

4. System Implementation and Computational Framework

The methodology described in Section 3 was implemented as a modular computational system to demonstrate how the predictive model and optimization procedure can be organized into an operational decision-support workflow. The implementation separates data preparation, predictive modelling, optimization, and monitoring functions while allowing the individual components to be executed sequentially.

4.1 System Architecture

The implemented system consists of four functional modules: data processing, predictive modelling, optimization, and performance monitoring.

The data processing module receives the industrial process records and prepares the variables required by the predictive model while maintaining the temporal structure of the observations.

The predictive modelling module loads the trained Random Forest Regression model and generates process-performance predictions for given operating conditions. The trained model provides the computational representation of the observed process used during optimization.

The optimization module receives the predictive model and evaluates candidate operating configurations generated by the Genetic Algorithm. The best feasible configuration identified by the optimization procedure is returned as the recommended operating condition.

The performance monitoring module records model predictions, optimization progress, convergence behaviour, and computational execution time. This provides information that can be used to assess the computational feasibility of the workflow.

The modular structure allows individual components to be updated independently without redesigning the complete system.

4.2 Computational Execution Flow

The implemented approach uses the sequential execution path starting from process data to operational recommendation.

In the first step, verified industrial observations are supplied into the computational pipeline. Then, the trained Random Forest model is employed to make predictions of the process performance under candidate operating conditions.

Then, the optimization part suggests and evaluates possible configurations using the predictive model. Configurations which fit the prescribed operational ranges are kept in the optimization process, and the best one is found in terms of the objective function given in Section 3.

After performing optimization, the obtained configuration is compared with the initial operating condition to measure the estimated gain in the process performance. The optimization history and computational execution details are also saved for further use.

Hence, Section 4 is the implementation part of the methodology, but not an independent analytical technique.

4.3 Computational Environment and Deployment Considerations

The computational approach was coded in Python 3.11 employing Scikit-learn 1.5.1, NumPy 2.1.0, and Pandas 2.2.2 libraries for data processing and prediction modelling. The visualisation and analysis were performed with Matplotlib and Plotly. The Genetic Algorithm was coded using custom optimisation library in combination with trained Random Forest model.

All experiments were conducted using a computer with Intel Core i7 processor and 16 GB RAM. The observed optimisation execution time was about 5.50 seconds per iteration, which gives an approximation of the computation power required for the proposed computational approach.

The implementation was evaluated using industrial historical observations rather than by connecting the computational approach directly to the industrial control system. Thus, the provided runtime estimation should be considered as an example of computational feasibility, not as a proof of its industrial applicability. In a real-life production setup, the computational approach can be connected to an industrial data acquisition and control systems, such as SCADA or DCS, depending on the requirements.

5. Results and Discussion

5.1 Dataset Quality and Experimental Overview

Following preprocessing and data validation, 1,503 valid observations were retained from the original 1,510 industrial spray-drying records. The preprocessing procedure removed incomplete or inconsistent observations while preserving the temporal structure of the operational data. The final dataset contained seven process variables representing temporal, thermal, environmental, and mechanical characteristics of the spray-drying process.

Table 1: Dataset Summary and Quality Assessment

Attribute	**Value**
Original observations	1,510
Valid observations	1,503
Number of process variables	7
Data retention	99.54%

Table 1 shows that 1,503 of the original 1,510 observations were retained, corresponding to a data retention rate of 99.54%. Thus, only seven observations were excluded during preprocessing, indicating limited information loss. The resulting dataset provided a sufficiently complete basis for the subsequent predictive modelling and optimization analyses.

5.2 Predictive Performance

The predictive performance of the Random Forest Regression (RFR) model was evaluated using (R^2), MAE, and RMSE for both the training and testing datasets.

Table 2: Random Forest Performance Metrics

Metric	Training	Testing	Time-Series Cross-Validation
R^2	0.8885	0.9144	0.8427 ± 0.2027
MAE (°C)	2.01	3.68	—
RMSE (°C)	10.13	7.77	—

From Table 2, the RFR model was able to achieve an R^2 of 0.8885 in the training data set and 0.9144 in the testing data set. From the R^2 of the testing data set, we can observe that about 91.44% of the variance in the target process variable was explained by the model. The MAE and RMSE in the testing data set were 3.68 °C and 7.77 °C respectively. The time-series cross-validation yielded an average R^2 of 0.8427 ± 0.2027. This demonstrates that there is relatively stable prediction performance over the time-series validation folds, although variations among the folds indicate varying conditions at the process level.

The testing R^2 was slightly better than the training R^2. However, this does not necessarily imply any form of over-fitting since the two data sets are different in terms of variability and composition.

Figure 1 below depicts the correlation between the real and predicted output temperature values.

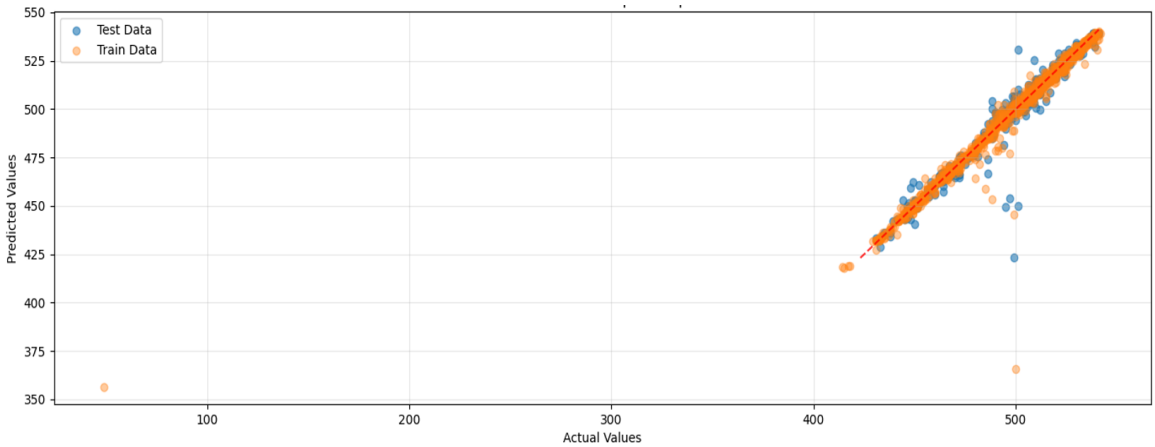

Figure 1: Actual versus Predicted Output Temperature Values

As illustrated in Figure 1, the predicted values generally follow the distribution of the observed output temperatures, with the majority of observations concentrated around the expected prediction trend. The visual agreement between observed and predicted values is consistent with the quantitative performance reported in Table 2 and provides graphical support for the model's predictive capability.

5.3 Temporal Prediction Performance

Because the observations represent industrial process measurements collected over time, the temporal behaviour of the predictions was examined in addition to the aggregate performance metrics.

Figure 2 illustrates the temporal relationship between the actual and predicted output-temperature values across the operational sequence.

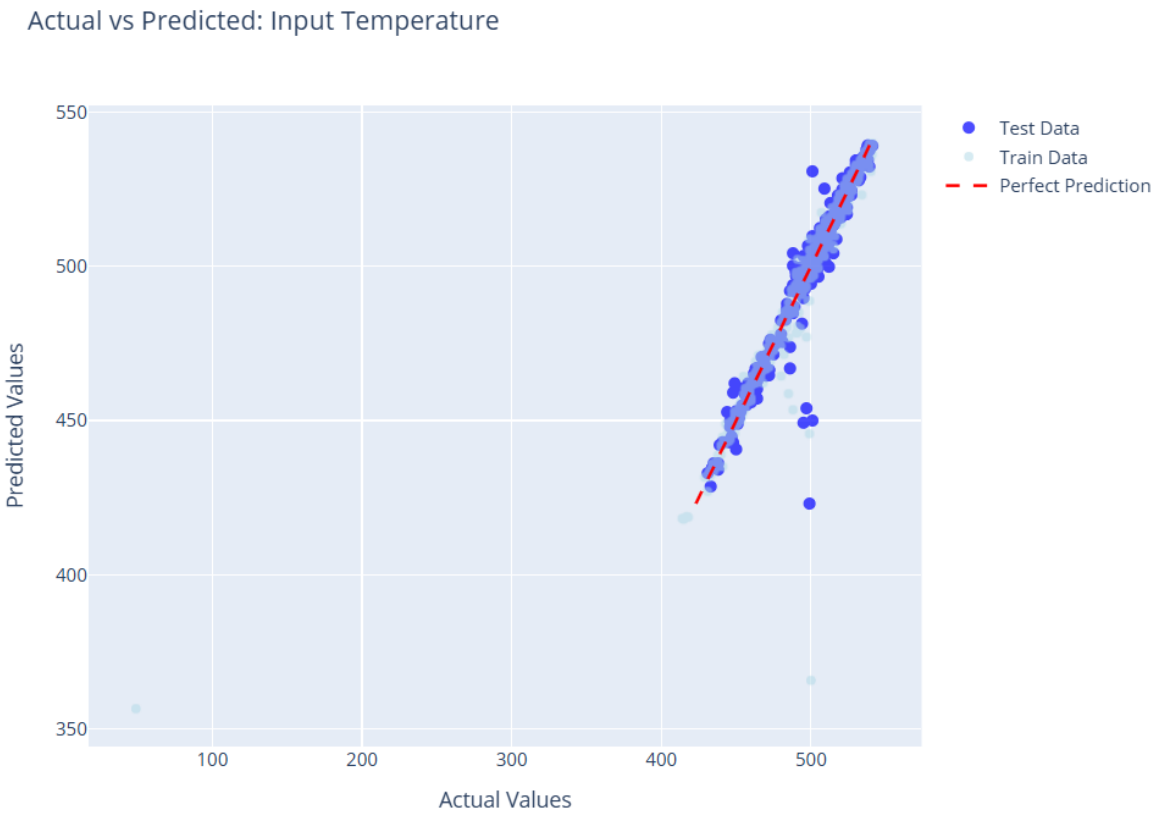

Figure 2: Temporal Comparison of Actual and Predicted Output Temperature Values

As illustrated in Figure 2, the predicted series is in line with the temporal pattern of the measured data. The model correctly predicts the significant variations and general trend of the output temperature, although the discrepancies can be noted during certain periods of high process variability. The discrepancies are anticipated in the real industrial setting in which operating conditions can change over time.

The temporal correspondence is supplemented by the overall statistics provided in Table 2 through demonstrating that the behavior of the model in terms of its predictions is not confined to just a few data points. The results obtained offer empirical confirmation of RQ1 about the modeling capabilities of RFR.

5.4 Feature-Importance Analysis

Feature importance was examined to determine which of the seven process variables contributed most strongly to the predictive behaviour of the RFR model.

Table 3. Random Forest Importance

Rank	Feature	Importance
1	Gas	0.7623
2	Time	0.1477
3	Outside Temperature	0.0650
4	Main Blade	0.0092
5	Humidity	0.0079
6	Tower	0.0068
7	Pressure	0.0057

It can be seen from Table 3 that Gas is the most important feature in the Random Forest classifier model, whose importance is estimated to be 0.7623. The next two most important features are Time (0.1477) and Outside Temperature (0.0650), while all other variables have relatively small reported importance scores.

The large importance score of the Gas variable shows that this variable made a significant contribution to the behavior of the model under consideration. Also, the importance of the Time variable shows that time data in the operational logs played an important role in making predictions. However, one should keep in mind that these importance scores reflect associations between the model and variables but not causal relations; thus, changing a variable would not necessarily result in a change in the target process.

The results obtained are directly linked to RQ3.

5.5 Genetic Algorithm Optimization Results

The Genetic Algorithm was evaluated according to the progression of its best fitness value over successive generations.

Table 4: Genetic Algorithm Convergence

Generation	Best Fitness
0	496.5227
10	499.3937
20	510.3173
30	519.1194
40	522.7804
50	524.7828
60	525.7141
70	525.8377
80	525.8726
90	525.8749
100	525.8759
110	525.8760
120	525.8760

Table 4 shows a progressive increase in the best fitness value from 496.5227 at generation 0 to 525.8760 at generation 120. The improvement was particularly pronounced during the first 50 generations, after which the rate of improvement progressively decreased. By approximately generation 80, the fitness value had approached the final solution, with only marginal changes occurring during the remaining generations.

The final fitness value represents an improvement of approximately 5.91% relative to the initial fitness value. The stabilization observed during the later generations indicates that the evolutionary search had approached a stable solution within the specified search space.

Because the optimization values represent the model-defined fitness objective, rather than directly observed physical measurements, the 5.91% improvement should be interpreted as an improvement in the optimization objective rather than as a statistically verified physical efficiency improvement.

5.6 Optimal Operating Configuration

The final GA solution identified an operating configuration associated with the highest fitness value obtained during the optimization process.

Table 5: Optimized Operating Configuration

Parameter	Optimized Value	Operational Interpretation
Gas Flow Rate	41	Stable thermal input
Output Temperature	94 °C	Model-identified operating condition
Humidity	6.24%	Controlled moisture condition
Pressure	20	Stable process pressure

Table 5 shows the operating configuration associated with the final optimized solution. The configuration includes a gas flow rate of 41, output temperature of 94 °C, humidity of 6.24%, and pressure of 20. Collectively, these values represent the parameter combination selected by the GA within the predefined operational search space.

The result demonstrates how the predictive model was translated into an operational recommendation. However, because the configuration was identified through the surrogate model, it should be considered a model-derived recommendation rather than experimentally verified evidence of improved physical process performance.

5.7 Discussion

Taken together, the results demonstrate a sequential relationship between prediction, interpretation, and optimization. Table 2 and Figures 1–2 show that the Random Forest model provided strong predictive

performance, while Table 3 identifies the variables contributing most strongly to the model's predictions. Table 4 demonstrates progressive improvement and convergence of the Genetic Algorithm, and Table 5 translates the final optimization result into a specific operating configuration.

The results therefore address the four research questions in a logically connected manner. RQ1 is addressed through the predictive-performance and temporal analyses. RQ2 is addressed through the GA convergence and optimized configuration. RQ3 is addressed through the feature-importance analysis. RQ4 is addressed by the overall workflow linking industrial observations, prediction, interpretation, and constrained optimization.

The findings should nevertheless be interpreted within the scope of the available data. The analysis is based on a single industrial spray-drying system, and the optimized configuration has not been independently validated through new physical experiments. Accordingly, the results demonstrate the analytical and computational feasibility of the framework rather than establishing universal process-optimality.

This distinction is important because the contribution of the study is not the introduction of a new Random Forest or Genetic Algorithm methodology. Rather, the results demonstrate how an established predictive model can be used with evolutionary search to move from historical industrial observations toward interpretable, constraint-based operational recommendations.

6. Summary, Conclusion, and Recommendations

6.1 Summary

This study developed and evaluated a data-driven framework for modelling and optimizing an industrial spray-drying process using Random Forest Regression (RFR) and a Genetic Algorithm (GA). The study focused on the practical integration of predictive modelling, model interpretation, and constrained optimization using historical industrial process data.

Following preprocessing and validation, 1,503 valid observations were retained from the original 1,510 records. The Random Forest model demonstrated strong predictive performance, achieving a testing R^2 of 0.9144, with a testing MAE of 3.68 °C and RMSE of 7.77 °C. The cross-validation results further provided evidence of predictive stability across the evaluation folds.

The feature-importance analysis indicated that Gas was the most influential predictor, followed by Time and Outside Temperature. These results provide an interpretable indication of the variables most strongly associated with the model's predictions and complement the predictive performance assessment.

The trained Random Forest model was subsequently used as a surrogate objective evaluator within the Genetic Algorithm. The optimization process improved the best objective-function value from 496.5227 at initialization to 525.8760 after 120 generations, corresponding to an improvement of approximately 5.91%. The optimization converged progressively, with most of the improvement occurring during the earlier generations and only marginal changes during the later stages.

Overall, the findings demonstrate the feasibility of using a Random Forest surrogate together with Genetic Algorithm search to connect industrial process prediction with constrained operational optimization. The contribution of the study therefore lies in the empirical evaluation of this workflow using real industrial spray-drying data rather than in proposing a new machine-learning or evolutionary optimization algorithm.

6.2 Conclusion

The findings indicate that Random Forest Regression can provide an effective predictive representation of the nonlinear relationships present in the investigated industrial spray-drying process. The testing R^2 of 0.9144 and the associated error measures demonstrate that the model was capable of explaining a substantial proportion of the observed variation in the target process variable.

The feature-importance results further indicate that the predictive behaviour of the model was dominated by a small number of process variables, particularly Gas, Time, and Outside Temperature. This provides an additional basis for interpreting the model and examining the operational conditions associated with its predictions.

The Genetic Algorithm subsequently used the trained Random Forest model to search the predefined operational space. The observed increase in the objective-function value from 496.5227 to 525.8760 demonstrates that the surrogate-based optimization procedure was able to identify a candidate operating configuration associated with improved predicted performance. The optimization should therefore be interpreted as a model-based recommendation, rather than as direct experimental confirmation of a physical process improvement.

Importantly, the study does not claim that Random Forest Regression or Genetic Algorithms are universally superior to alternative machine-learning or optimization techniques. Instead, the results demonstrate that the selected methods can provide a computationally practical workflow for the specific industrial dataset and spray-drying process investigated.

The principal scientific contribution is consequently the empirical demonstration of a pathway from **real industrial process data → predictive modelling → feature-importance analysis → constrained evolutionary search → candidate operational recommendation**. This provides a practical basis for data-driven decision support in industrial spray-drying applications while recognizing the need for further validation before direct deployment in live production environments.

6.3 Recommendations

Based on the findings and limitations of the present study, the following recommendations are proposed.

6.3.1 Industrial Recommendations

1. Use predictive models as decision-support tools: Random Forest Regression can be considered as a practical predictive component for industrial process monitoring and decision support where nonlinear relationships exist among operational variables.
2. Incorporate model interpretation into operational analysis: Feature-importance information can assist engineers and operators in identifying variables that contribute most strongly to model predictions and in evaluating the plausibility of optimization recommendations.
3. Apply optimization within predefined operational constraints: Evolutionary optimization should be implemented within physically and operationally feasible parameter ranges so that model-generated recommendations remain consistent with actual process requirements.
4. Validate optimized recommendations experimentally before deployment: The optimized configuration identified by the computational framework should be subjected to controlled industrial trials before being used for routine process control. The optimization results presented in this study represent model-based predictions and should not be interpreted as experimentally verified process improvements.

6.3.2 Research Recommendations

1. Analyzing alternative prediction models: In future studies, there can be comparisons between Random Forest Regression and competing techniques such as XGBoost, LightGBM, CatBoost and other predictive models to investigate whether choice of model has any effect on the optimization results.
2. Analyzing alternative optimization algorithms: In future studies, there can be comparisons between Genetic Algorithms and methods like Particle Swarm Optimization and Differential Evolution or other metaheuristic techniques under the same constraints.
3. Conducting prospective industrial validations: Future studies must conduct optimizations of the operation configurations through physical experiments or prospective production tests to validate whether there is improvement in actual process settings.
4. Conducting tests on extended data sets or industrial environments: Conducting tests on other production spans, or other spray-dryers or industrial processes will help understand the applicability of the proposed framework in other settings.
5. Using better interpretability and uncertainty quantification: Future research can utilize better explainability and uncertainty quantification techniques for better understanding of prediction and optimization performance.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning, 109*(11), 1997–2028. https://doi.org/10.1007/s10994-020-05910-7

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation, 6*(2), 182–197. https://doi.org/10.1109/4235.996017

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.*

Farinola, L. A. (2026). *A framework for implementing Industry 4.0 technologies in smart manufacturing environments.* Kocaeli Journal of Science and Engineering. Accepted for publication.

Farinola, L. A., & Bazarkhan, D. (2025). Optimization of Complex Spray Drying Operations in Manufacturing Using Machine Learning: Evaluating Techniques for Energy Efficiency and Product Quality Enhancement. Open Journal of Applied Sciences, 15(9), 2662-2691. https://www.scirp.org/journal/paperinformation?paperid=145657

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

Holland, J. H. (1975). *Adaptation in natural and artificial systems.* University of Michigan Press.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Kagermann, H., Wahlster, W., & Helbig, J. (2013). *Recommendations for implementing the strategic initiative INDUSTRIE 4.0: Securing the future of German manufacturing industry.* Forschungsunion/acatech. The official acatech record confirms the 2013 final report.

Lee, J., Bagheri, B., & Kao, H.-A. (2015). A cyber-physical systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters, 3*, 18–23. https://doi.org/10.1016/j.mfglet.2014.12.001

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774.

Pham, D. T., & Afify, A. A. (2005). Machine learning in automated manufacturing. *Journal of Intelligent Manufacturing, 16*(3), 307–314. https://doi.org/10.1007/s10845-005-7026-3

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Vogel-Heuser, B., Fay, A., Schaefer, I., & Tichy, M. (2018). Evolution of software in automated production systems: Challenges and research directions. *Journal of Systems and Software, 143*, 1–16. https://doi.org/10.1016/j.jss.2018.05.011

Wahab, W., Alshamsi, R., Alharsousi, B., Alnuaimi, M., Alhammadi, Z., & Al-Zaitone, B. (2025). Recent Developments in Pharmaceutical Spray Drying: Modeling, Process Optimization, and Emerging Trends with Machine Learning. *Pharmaceutics, 17*(12), 1605. https://doi.org/10.3390/pharmaceutics17121605

Wuest, T., Weimer, D., Irgens, C., & Thoben, K.-D. (2019). Machine learning in manufacturing: Advantages, challenges, and applications. *Production & Manufacturing Research, 7*(1), 1–23. https://doi.org/10.1080/21693277.2018.1522620

Yin, S., Li, X., Gao, H., & Kaynak, O. (2015). Data-based techniques focused on modern industry: An overview. *IEEE Transactions on Industrial Electronics, 62*(1), 657–667. https://doi.org/10.1109/TIE.2014.2308133