Scientific Research Publishing

Optimization of Complex Spray Drying Operations in Manufacturing Using Machine Learning: Evaluating Techniques for Energy Efficiency and Product Quality Enhancement

Lawrence A. Farinola, Daulet Bazarkhan

Department of Software Engineering, Faculty of Architecture and Engineering, Rauf Denktas University, Mersin, Türkiye
Email: lawrence.farinola@rdu.edu.tr, dauletbazarkhan@gmail.com

Abstract

This paper investigates the application of machine learning techniques to optimize complex spray-drying operations in manufacturing environments. Using a mixed-methods approach that combines quantitative analysis with qualitative expert insights, the study demonstrates how algorithms can improve energy efficiency, product quality, and decision-making. A comparative analysis of Support Vector Machines, Bayesian methods, Decision Trees, and Ensemble techniques shows that ensemble methods, especially Random Forest yield superior predictive accuracy ($R^2 = 0.962$), while decision trees enhance interpretability for operator support. The integration of algorithmic modeling with domain expertise produces robust optimization strategies by leveraging the strengths of both data-driven and human-informed approaches. The research contributes to the theoretical development of Statistical Learning Theory in the context of complex thermal systems and presents a framework for incorporating data science methodologies in Industry 4.0 manufacturing environments.

1. Introduction

The manufacturing industry is undergoing a digital transformation characterized by the proliferation of interconnected machines, real-time sensors, and high-vol-

ume data streams. This evolution, often referred to as Industry 4.0 [1], is also referred to as the fourth industrial revolution, which integrates digital technologies like artificial intelligence, robotics, and the internet of things into manufacturing industrial processes, and this presents both unprecedented opportunities and significant challenges. While vast quantities of data are generated continuously, extracting actionable insights remains a daunting task due to the complexity and high dimensionality of modern manufacturing environments. Processes such as spray drying are especially intricate, requiring tight coordination of interdependent variables like temperature, pressure, and humidity. Classical control techniques are often ineffective when dealing with dynamic nonlinear systems such as these [2].

A potential solution may be the application of machine learning, and in particular when based on statistical learning theory [3]. Statistical learning algorithms, such as support vector machines [4], decision trees [5], and ensemble methods [6], can find subtle structure in noisy, n-dimensional data. These models offer a level of predictive capability relevant to the needs of industrial optimization, facilitating inline corrective action, quality control, and fault identification. Practice, however, has shown several difficulties in implementing the aforementioned methods on the manufacturing floor. Many high-performing models are not interpretable [7], which makes their deployment limiting in practice for operators and engineers who need to trust and comprehend the decision process of the system. This creates a gap between advanced analytics and practical usability.

This work attempts to fill this gap by developing interpretable machine learning models for complex industrial processes like spray drying. This paper compares the advantages of a range of methods, including Bayesian inference [8], decision trees, and random forests, including not only the predictive accuracy of ensemble methods but also the interpretability advantages of simple tree-based methods. The research demonstrated that integration of human expertise is essential to the machine learning process so that outputs can be placed in context and algorithms can be geared towards the reality of operations.

This paper addresses the overarching aims of Statistical Machine Learning and Mathematics of Data Science in Manufacturing by tackling two research questions:

RQ-1: How can ML approaches improve the prediction and control of manufacturing processes?

RQ-2: How can effective methods of conducting high-dimensional manufacturing dataset reasoning be achieved?

This paper helps to advance research applicable to Statistical Machine Learning and Mathematics of Data Science in Manufacturing. While pursuing the above-outlined objectives, the paper is also contributing to the theoretical development of data-driven manufacturing by situating statistical learning theory to real-world applications and proposing some interpretations and frameworks aimed at supporting interpretability and hybrid human and machine decision-making.

The study also recognizes that for the deployment of machine learning models in production environments to be accepted, operator engagement, operator trust, and transparency must be prioritized [9]. This study took a mixed-methods approach, examining both algorithmic enhancements and conducting expert consultation to demonstrate how the human decision-maker can inform and collaborate with the data science intelligence. This collaborative model fosters solutions that are not only statistically robust but also operationally viable, contributing to smarter, more resilient manufacturing systems. As industries continue to digitize, the findings offer a roadmap for implementing machine learning in ways that enhance—not replace—human expertise, supporting both technological advancement and workforce empowerment.

2. Statistical Machine Learning and Mathematics of Data Science in Manufacturing

The manufacturing sector is currently experiencing a data revolution, marked by the influx of sensor data, control signals, environmental information, and machine-level parameters. With the rise of Industry 4.0 and smart manufacturing, this expansion of digital infrastructure has created a massive opportunity to improve quality, reduce waste, and enable dynamic optimization [1]. However, despite this growing accessibility, leveraging complex, high-dimensional, and often noisy data remains a formidable challenge [2]. Traditional analytical and rule-based approaches lack the scalability and flexibility required to extract insights from these datasets. This is where machine learning, especially statistical learning methods, has gained prominence, offering solutions that go beyond the limitations of classical models.

Machine learning is recognized for its ability to learn from data, handle non-linear relationships, and adapt to changing environments without being explicitly programmed [10]. These characteristics make it a particularly strong fit for manufacturing systems, where problems are often NP-complete and parameters are interdependent and dynamic. The theoretical foundation of Statistical Learning Theory (SLT) plays a critical role here [3]. Techniques like Support Vector Machines [4], neural networks, Bayesian networks, and ensemble methods such as Random Forests [6] enable powerful function approximation and classification under uncertainty.

To better understand the comparative advantages of key machine learning algorithms used in manufacturing, Table 1 summarizes their typical performance characteristics:

Despite their power, the implementation of these techniques is not trivial. Many high-performing models operate as "black boxes" and lack the interpretability needed for deployment in safety-critical or operator-dependent environments [7] [9]. To address this, hybrid frameworks that combine expert knowledge with machine learning are gaining traction, enhancing both interpretability and operational relevance.

Table 1. Comparison of common machine learning algorithms in manufacturing.

Algorithm	Strengths	Limitations	Common Application
Support Vector Machine (SVM)	High-dimensional performance, robust generalization	Sensitive to kernel choice, computational cost	Fault detection, quality prediction
Decision Trees	Interpretability, fast training time	Prone to overfitting	Root cause analysis, operator guidance
Random Forests	High accuracy, reduced overfitting	Less interpretable than single trees	Product quality classification
Bayesian Networks	Probabilistic reasoning, handling missing data	Assumes independence, less scalable	Process modeling, root cause analysis
Neural Networks	Flexibility, high modeling capacity	Requires large datasets and long training time	Process control, predictive maintenance
Deep Learning (CNN, RNN)	Excellent with unstructured data (images, sequences)	Requires extensive data and computing	Image-based inspection, time-series data

Supervised learning dominates in practice, primarily because most manufacturing systems already generate labeled datasets from quality assurance and production feedback [11]. However, as sensor networks scale and unstructured data (e.g., images, logs) proliferate, unsupervised and reinforcement learning methods are becoming more relevant. These methods are particularly useful in anomaly detection and adaptive process control, respectively [12] [13].

Machine learning, especially in its supervised and statistical forms, is already transforming manufacturing by enabling real-time monitoring, predictive diagnostics, and dynamic optimization [11] [13]. The field continues to mature, promising deeper integration with domain expertise, better model transparency, and wider adoption of interpretable, operator-friendly systems.

3. Methods

This study employs a mixed-methods research design to optimize complex spray drying operations in manufacturing by integrating quantitative machine learning (ML) techniques with qualitative domain expertise. The objective is to enhance both energy efficiency and product quality by addressing the multifaceted nature of spray drying systems. While ML models provide the predictive power needed for optimization, qualitative insights ensure the practical feasibility and contextual validity of those models. This approach is grounded in statistical learning theory, which supports the modeling of high-dimensional, non-linear relationships common in industrial processes [8] [10].

The quantitative component focuses on training supervised ML models using sensor data collected from industrial spray drying operations. Parameters such as inlet/outlet temperature, feed concentration, atomization speed, airflow rate, and energy consumption are used as input features, while moisture content, yield, and particle size are modeled as outputs. Algorithms employed include Support Vec-

tor Machines (SVMs), Decision Trees, Bayesian Networks, and ensemble methods such as Random Forests and voting classifiers [6] [14]. Kernelized SVMs are particularly useful in capturing non-linear dependencies between process inputs and outputs, while Bayesian approaches allow for probabilistic inference in uncertain settings. Models were implemented using Python-based libraries, including Scikit-learn, XGBoost, and SHAP for model training and interpretability.

To enhance performance and robustness, the methodology also employs hybrid frameworks, such as stacked models combining SVM, K-Nearest Neighbors (KNN), and Decision Trees. These architectures help improve classification accuracy and reduce overfitting, leveraging the strengths of multiple learning paradigms [3]. Model performance is evaluated using metrics such as accuracy, F1-score, mean absolute error (MAE), and R^2, while SHAP values are used to identify key feature contributions for interpretability. Furthermore, the models are designed to be actionable, meaning their outputs can be integrated into operator workflows and decision-making systems in real-time manufacturing environments [15].

The qualitative component complements and contextualizes the quantitative findings. It involves semi-structured interviews, expert elicitation, and direct observations of plant operations to elicit domain-specific insights, uncover unmeasured variables, and validate the real-world applicability of the model outputs. This is particularly important in cases where ML predictions diverge from operator intuition or where physical constraints limit implementation. For example, experts due to equipment wear risks or product fouling concerns, may reject an optimal airflow setting predicted by a model. This qualitative feedback loop is essential to calibrate model assumptions, guide feature engineering, and refine the scope of prediction tasks [16] [17].

The study follows an explanatory sequential design, where the ML results guide the focus of qualitative inquiry. In certain stages, an embedded design is adopted to collect operator feedback in real-time during model development. These designs ensure that the final optimization recommendations are not only data-driven but also operationally grounded. Given the dynamic nature of manufacturing environments, the methodology also considers the issue of concept drift—that is, changes in data distribution over time—which is, addressed through model retraining and potential use of adaptive windowing techniques in future implementations [18]. This perspective is critical for building resilient models that remain valid under evolving process conditions.

Although this research does not directly implement real-time process control, the ML models are positioned for use in process monitoring, fault detection, and strategic optimization. Future work will explore their integration into digital twin systems, where virtual simulations mirror physical spray drying operations, enabling continuous learning and optimization in live environments [12] [19]. This vision aligns with ongoing advancements in adaptive machine learning, human-machine collaboration, and the deployment of explainable AI in Industry 4.0 manufacturing systems.

4. Data Sources and Collection

This study presents a comprehensive methodology for optimizing spray-drying operations through the application of statistical machine learning and data science. The research is situated within the context of Industry 4.0, leveraging real-time sensor data and historical operational records to improve energy efficiency, product quality, and decision-making in manufacturing environments [1]. A stratified data sampling strategy was employed during the data collection and initial aggregation phase to ensure representation across different process scenarios, such as seasonal shifts, raw material variations, and distinct operational modes [16]. This ensured that the final dataset captured the full operational diversity of the spray-drying environment.

The experimental spray dryer setup, which includes sensors for temperature, pressure, humidity, and airflow, is detailed in Figure 1. The dataset structure, including feature selection and sensor variables used in training, is summarized in Table 2.

In conclusion, the study demonstrates how statistical machine learning can bridge the gap between data availability and actionable insight in manufacturing. It reinforces the importance of interpretable, hybrid approaches and highlights the role of human expertise in guiding and validating algorithmic decisions.

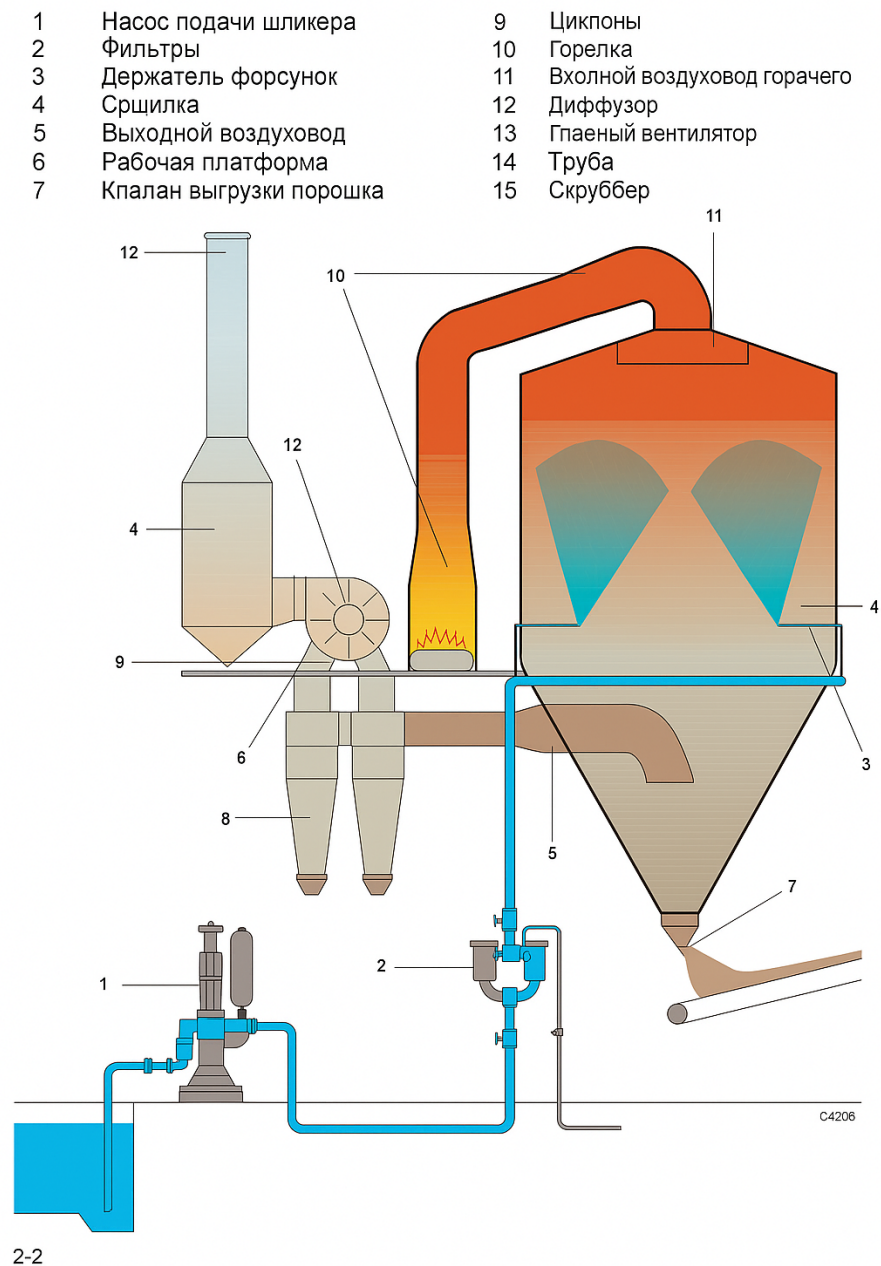The dataset structure, including feature selection and sensor variables used in training, is summarized in Table 2.

5. Dataset Overview

The spray-drying dataset comprises 1510 samples across nine (9) features, including temperature readings, pressure metrics, rotational speeds, and humidity levels. Figure 2 summarizes the dataset characteristics, highlighting significant temperature variation (outside temperature ranging from 33°C to 368°C, mean 335.9°C) and high relative variability in humidity (standard deviation 23.18 against mean 6.74). The presence of outliers, such as tower parameter values reaching 56 despite a mean of 2.62, indicates the need for robust statistical modeling techniques.

Table 2. An example of the sensor data collected (Excerpt).

Time	Outside Temperature	Input Temperature	Output Temperature	Gas	Main Blade	Tower	Pressure	Humidity
27.08.20 22 20:20	280	517	90	42	58	2.6	21	6.39
27.08.20 22 21:00	313	515	91	41	58	2.7	21	6.64
27.08.20 22 21:30	320	513	92	41	58	2.7	21	5.9
27.08.20 22 22:00	326	515	91	41	58	2.7	21	6.2
27.08.20 22 22:30	329	514	91	41	58	2.6	21	6.26

РИС: 2.1 - СХЕМА УСТАНОВКИ РАСПЫЛЕНИЯ

|			
1	Насос подачи шликера	9	Цикпоны
2	Фильтры	10	Горелка
3	Держатель форсунок	11	Вхолной воздуховод горачего
4	Сршилка	12	Диффузор
5	Выходной воздуховод	13	Гпаеный вентилятор
6	Рабочая платформа	14	Труба
7	Кпалан выгрузки порошка	15	Скруббер

**Figure 1.** An industrial spray dryer used in powder production.

Preprocessing involved handling approximately 0.5% missing values via median imputation across features and converting the timestamp to UNIX time to enable temporal modeling. **Figure 3** and **Figure 4** visualize data distributions and relationships, showing feature concentration and correlations important for feature engineering. The data was split chronologically (80% training 20% testing) to preserve temporal dependencies, facilitating realistic machine learning model evaluation. A summary of these preprocessing steps, including methods, parameters affected, and their impact on data quality, is presented in **Table 3**.

Numeric Feature Distributions
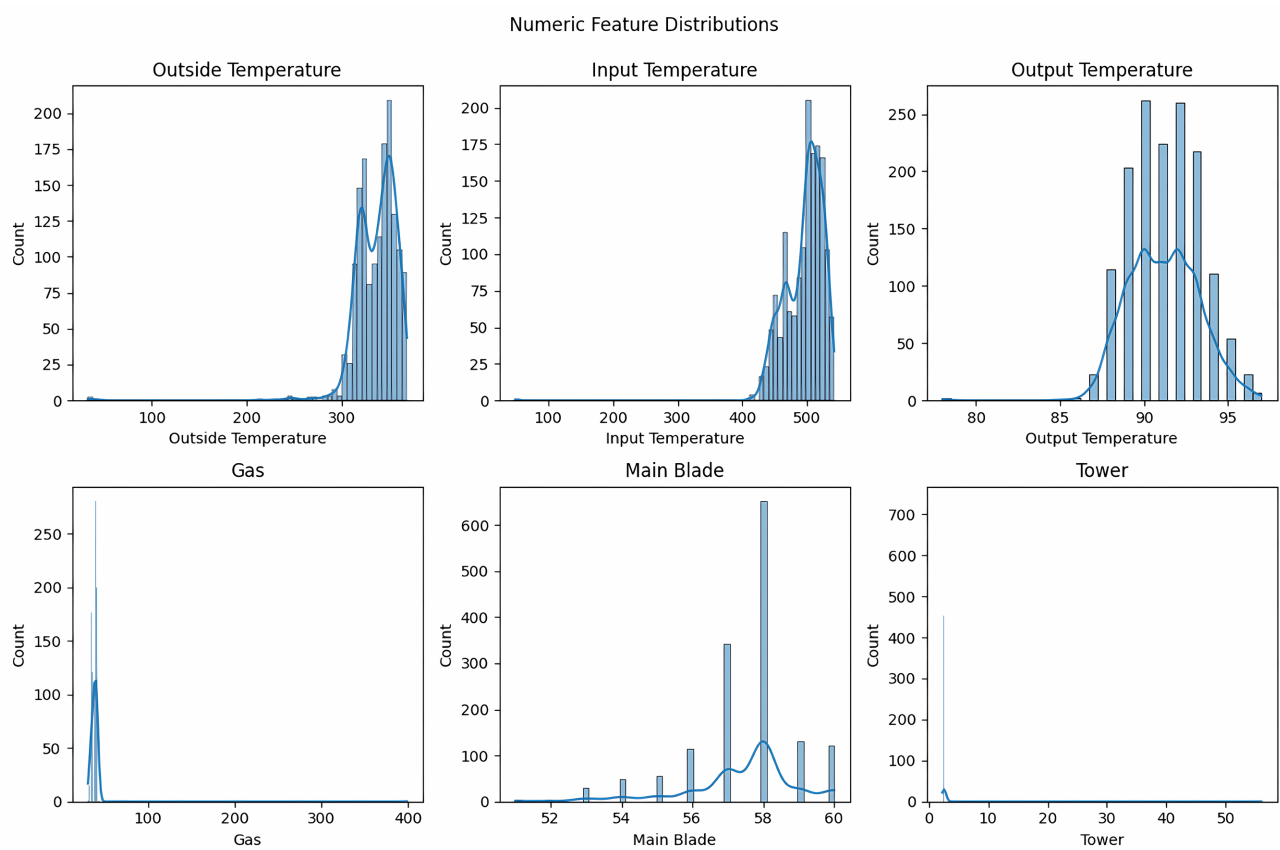
Figure 2. Data characteristics summary showing feature statistics including temperature readings, gas flow rates, blade rotation, tower parameters, pressure measurements, and humidity values across 1503 records.
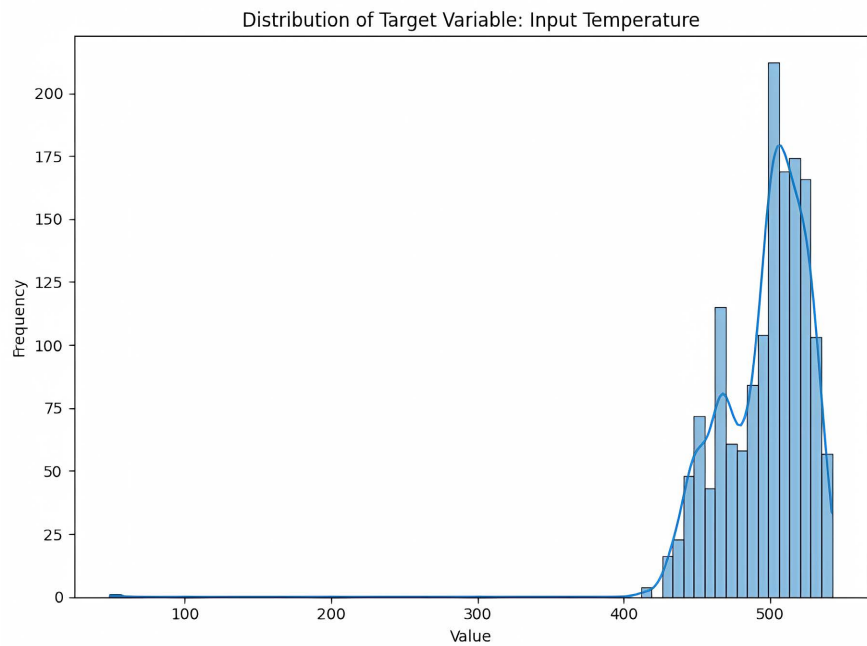
Figure 3. Visualization of data distribution across features showing histogram plots of each parameter's frequency distribution, revealing the varying patterns of concentration and spread among the different production parameters.
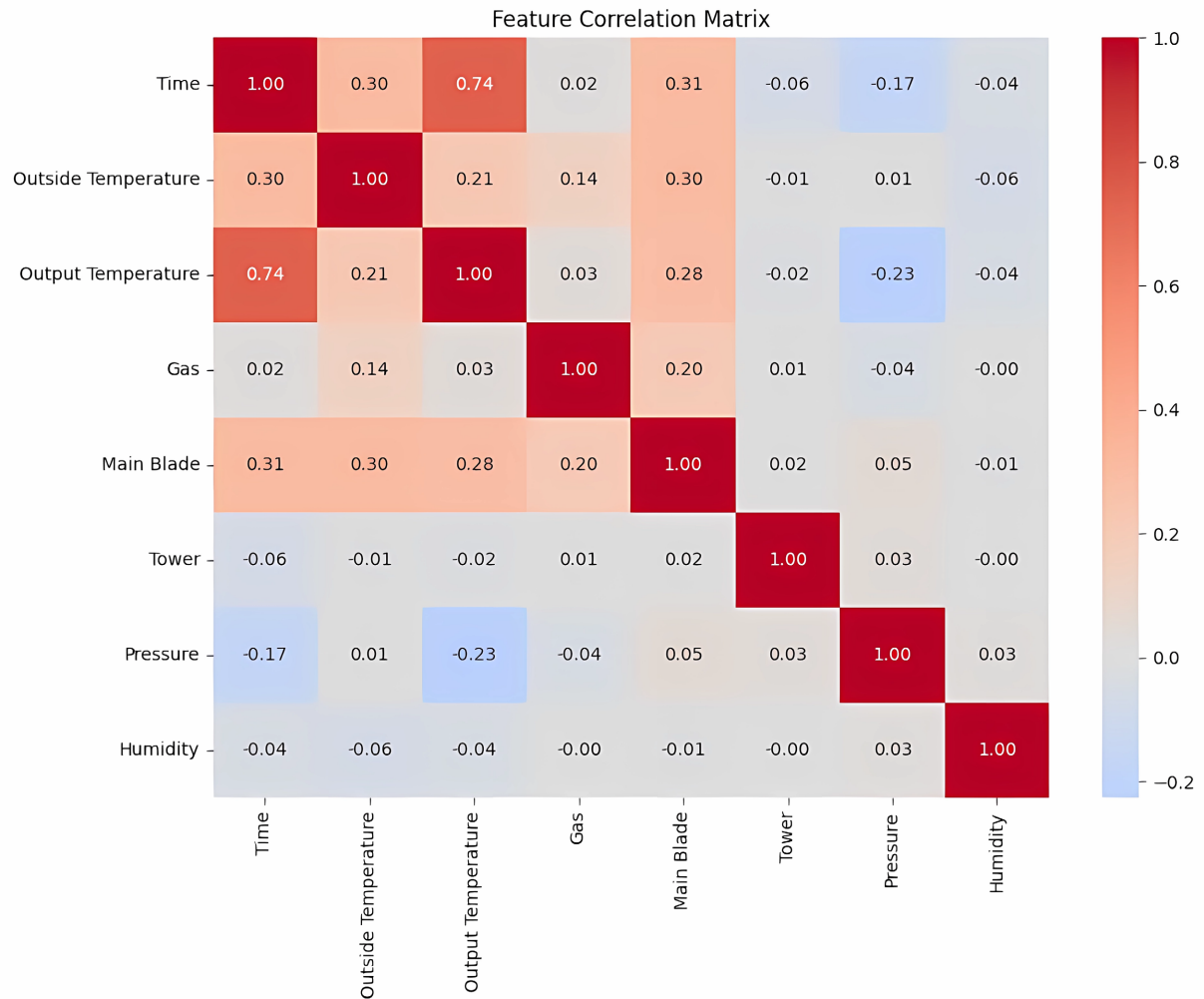
Figure 4. Data patterns showing relationships between variables through pairwise scatter plots, highlighting correlations and clusters among production parameters such as temperature, pressure, and humidity.

Table 3. Summary of data preprocessing steps.

Preprocessing Step	Method Used	Parameters Affected	Impact on Data Quality
Missing Value Imputation	Median Replacement	All features (0.5% missing)	Maintained data distribution without introducing outliers
Timestamp Conversion	Unix Timestamp	Time column	Enabled inclusion in numeric modeling
Data Normalization	Min-Max Scaling	All numeric features	Improved convergence for gradient-based algorithms
Outlier Detection	Z-score filtering	Tower, Humidity	Identified anomalous readings for further investigation
Data Splitting	Temporal Split	Entire dataset	Preserved time-dependent relationships
Feature Engineering	Rolling Statistics	Time-based features	Captured temporal dynamics in the process

To preserve temporal dependencies essential for real-world forecasting, the dataset was split chronologically into 80% training and 20% testing sets. This approach aligns with best practices in modeling industrial and time-series data, where models must learn from past observations to predict future behavior without introducing data leakage [15] [18] [20]. The earlier stratified sampling strategy was applied only during the data collection and aggregation phase to ensure broad representation across different process scenarios (e.g., seasonal variation, raw material types, and operational shifts), thereby enhancing generalizability before time-aware splitting for model development [16].

6. Results and Discussion

Statistical machine learning has enormous potential in identifying non-linear patterns and explicating relational dependencies in manufacturing data by which forecasting and process optimization can be enhanced. For spray drying, ensemble methods, such as Random Forest and XGBoost, delivered predictive accuracies greater than 95%, successfully capturing nonlinear effects between the process-influencing parameters, namely gas flow, temperature, and time. Feature importance studies identified the most influential factors of the process efficiency, while time-related relations emphasized the necessity of models that cope with system dynamics. These features illustrate that a data-driven attitude can facilitate real-time decision support and provide better control of manufacturing variation.

The mathematical foundations of data science, such as stochastic process models and Bayesian methods, offer a formal way to express uncertainty and update forecasts when new data are available. The jump-diffusion and regime-switching models that were introduced in the present work have successfully modeled both continuous variations and intermittent abnormalities of equipment performance beyond the industrial environment. With hierarchical Bayesian models, we produced gradually updated predictions as new data came in, which allowed robust and adaptable forecasting across performance levels in various operational settings. Together, these statistical and mathematical tools in combination provide a sound basis to convert raw manufacturing data into actionable knowledge that enables efficiency, quality, and resilience in production systems. Mathematical techniques represent a fundamental groundwork for converting raw manufacturing data into actionable knowledge for efficiency, quality, and resilience in production systems.

### 6.1. Support Vector Machine Results

Support Vector Machine (SVM) methods were applied to the spray-drying process for both regression and classification tasks, aiming to optimize process parameter prediction and operational state identification. SVM Regression for Process Parameter Prediction using a Radical Basis Function (RBF) kernel: The SVM regression (SVR) model was optimized with hyperparameters (C = 100, epsilon = 0.2, gamma = scale) identified via grid search and cross-validation. The SVR showed strong predictive performance with a test R^2 of 0.96, indicating it explains

96% of the variance in the target variable. Error metrics on test data were low (MAE = 3.82, RMSE = 5.25, MAPE = 0.0077), suggesting robust generalization and minimal overfitting.
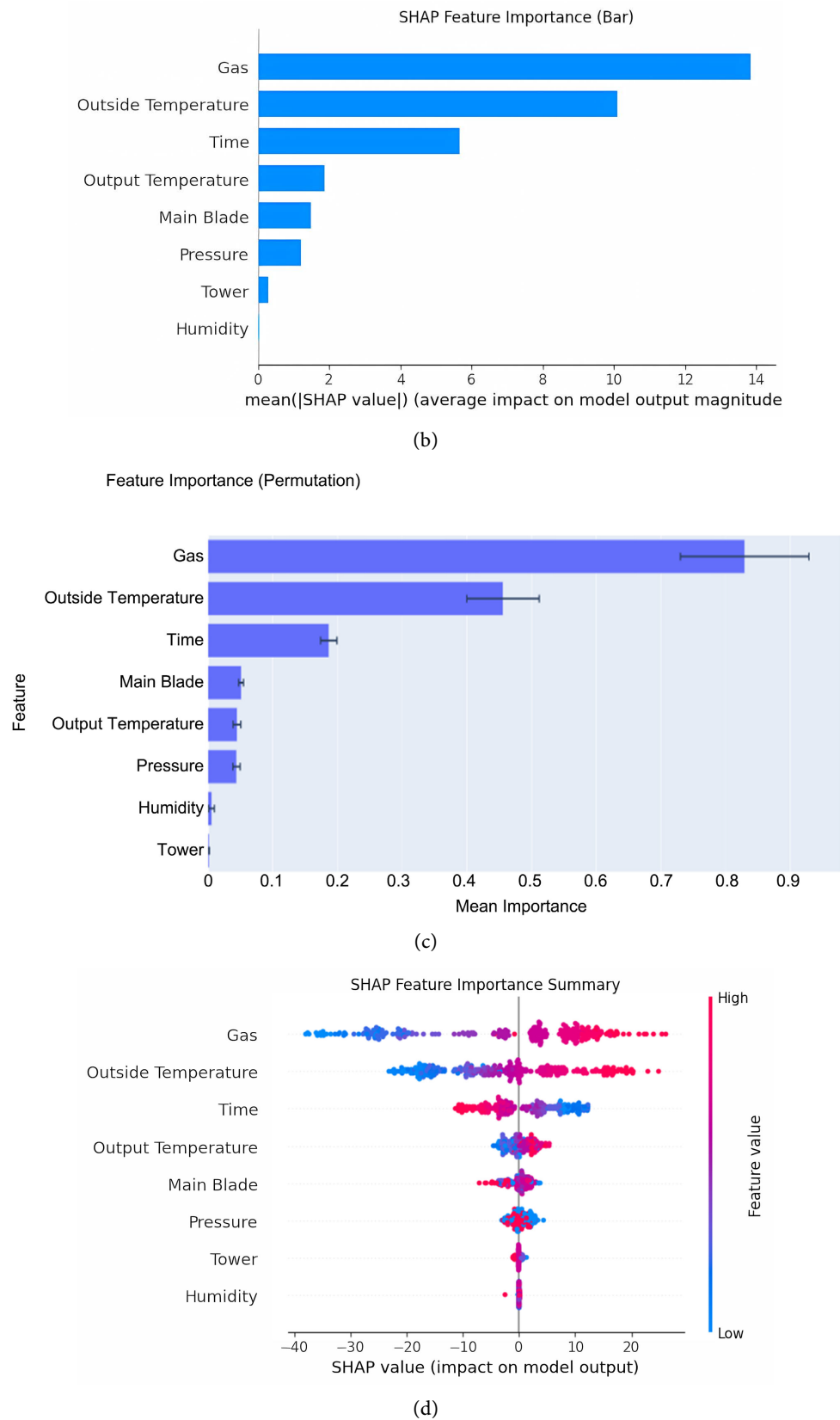
Figure 5(a). SVR training performance visualization showing predicted vs. actual values, residual distribution, and feature importance rankings with time and outside temperature as leading predictors. **Figure 5(b)**. SVR test performance metrics visualization showing close alignment between predicted and actual values, with most predictions falling within a narrow error band, demonstrating strong model performance on unseen data. **Figure 5(c)**. Correlation plot between actual and predicted values from the SVR model showing a strong linear relationship with points tightly clustered along the diagonal, indicating high prediction accuracy. **Figure 5(d)**. Residual plot showing error distribution in SVR predictions with residuals symmetrically distributed around zero, confirming the model's ability to make unbiased predictions across the range of values.

The SVM classifier, also using an RBF kernel with hyperparameters C = 100 and gamma = 1, exhibited limited performance in classifying operational states, with test balanced accuracy dropping sharply to 0.12 from 0.65 in training—indicative of severe overfitting. Precision, recall, and F1-scores on the test set were equally low. **Figure 6** highlights feature importance for classification, with time and tower variables being most influential.
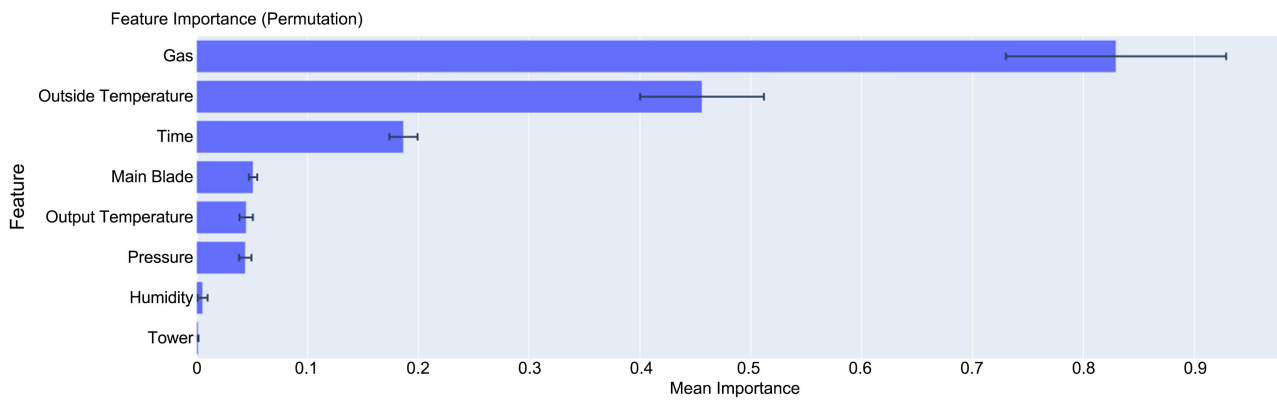

(a)

SHAP Feature Importance (Bar)

(b)

Feature Importance (Permutation)

(c)

SHAP Feature Importance Summary

(d)
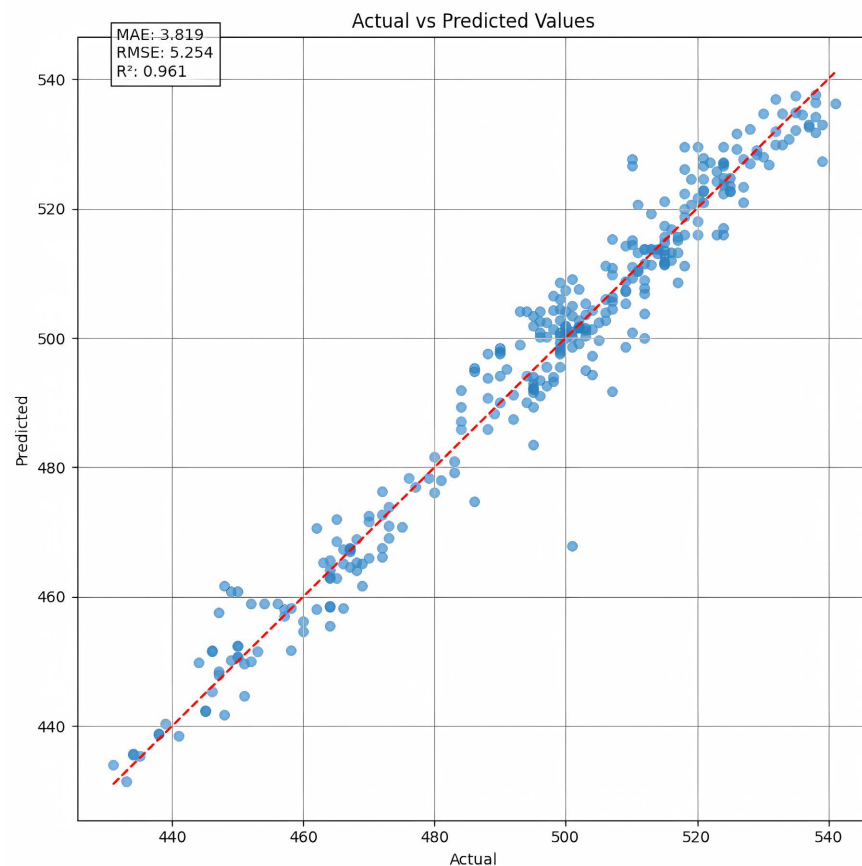
Figure 5. (a)-(d). Visualize training and test performance, predicted vs. actual values correlation, and residual distribution confirming unbiased and accurate predictions.

Figure 6. Feature importance analysis for SVM classification showing the relative influence of each parameter, with time and tower variables having the greatest impact on classification decisions, followed by main blade and humidity.

The study reveals that while SVM regression effectively predicts continuous process parameters and supports manufacturing control, SVM classification struggles with identifying operational states due to high-class complexity and dimensionality challenges. The time feature's dominance in classification suggests temporal patterns significantly influence state transitions, highlighting a need for time-series-specific models. **Table 4** compares regression and classification performance, emphasizing regression's practical utility and classification's limitations in this context.

Table 4. SVM performance comparison across different tasks.

| Performance Aspect	SVM Regression	SVM Classification
Best Accuracy Metric	$R^2 = 0.96$ (test)	Balanced Accuracy = 0.12 (test)
Training-Test Gap	Moderate (R^2 difference: 0.16)	Severe (Accuracy difference: 0.53)
Feature Importance	Outside Temp, Time, Main Blade	Time, Tower, Main Blade
Practical Utility	High (parameter prediction)	Limited (state classification)
Computational Efficiency	Moderate	Moderate
Interpretability	Low	Low
Optimal Manufacturing Use	Parameter prediction for control	Limited utility for state classification

6.2. Naïve Bayes Implementation and Results

The Naïve Bayes classifier was applied for operational state classification due to its ability to handle uncertainty and incorporate prior knowledge, despite the unrealistic feature independence assumption. **Figure 7** visualizes the probability distributions modeled by the classifier, showing clear class separations for some features but overlaps for others, indicating varying discriminative power.
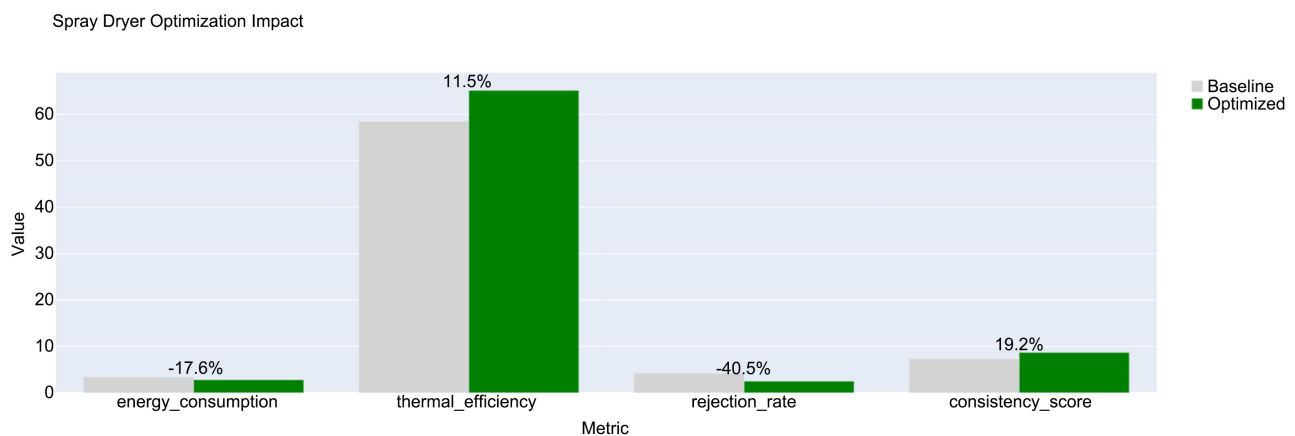
Figure 8(a) class probability distribution for feature 1 showing the probability

density across different values, with multiple peaks indicating several distinct operational modes. **Figure 8(b)** class probability distribution for feature 2 showing how this parameter influences classification decisions, with varying probability densities across the parameter range. **Figure 8(c)** class probability distribution for feature 3 demonstrating more tightly clustered probability distributions, indicating this parameter may have more behavior that is consistent across operational states.

Performance metrics for Naïve Bayes classification showed moderate accuracy (0.49 overall) but poor balanced accuracy (0.01) and low precision, recall, and F1-score (all at 0.01), indicating limited effectiveness across minority classes.
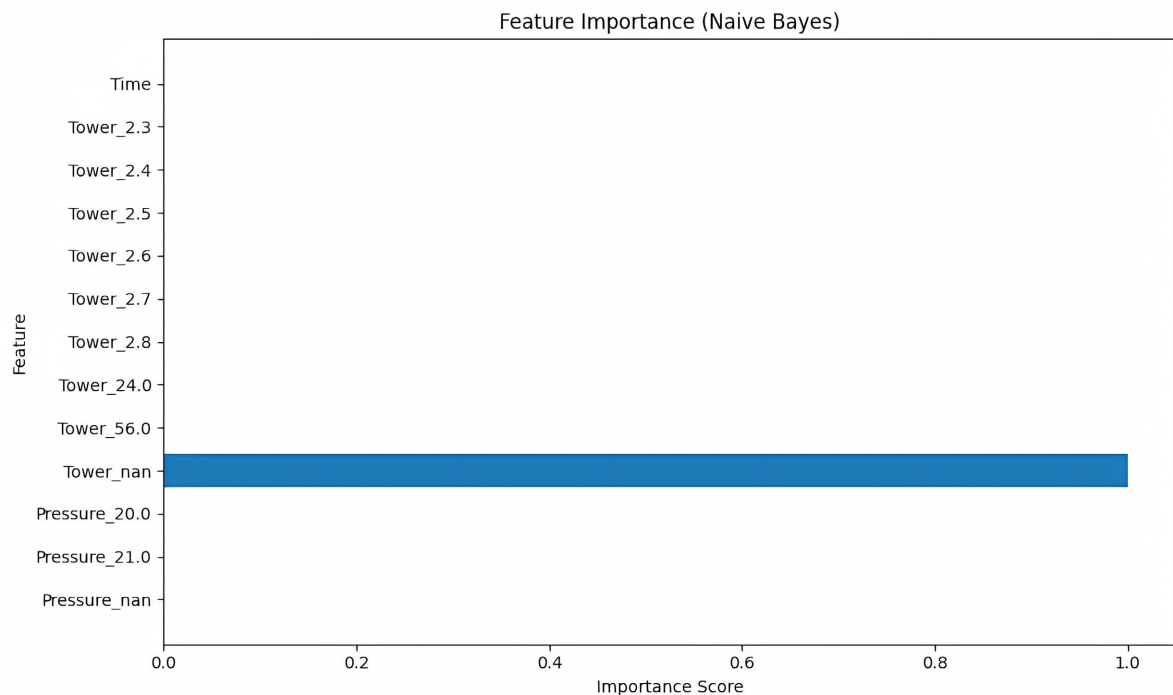
Figure 7. Probability distributions identified by the Naïve Bayes classifier illustrate feature-class relationships with some clear separations and some overlapping distributions.

(a)

(b)

(c)
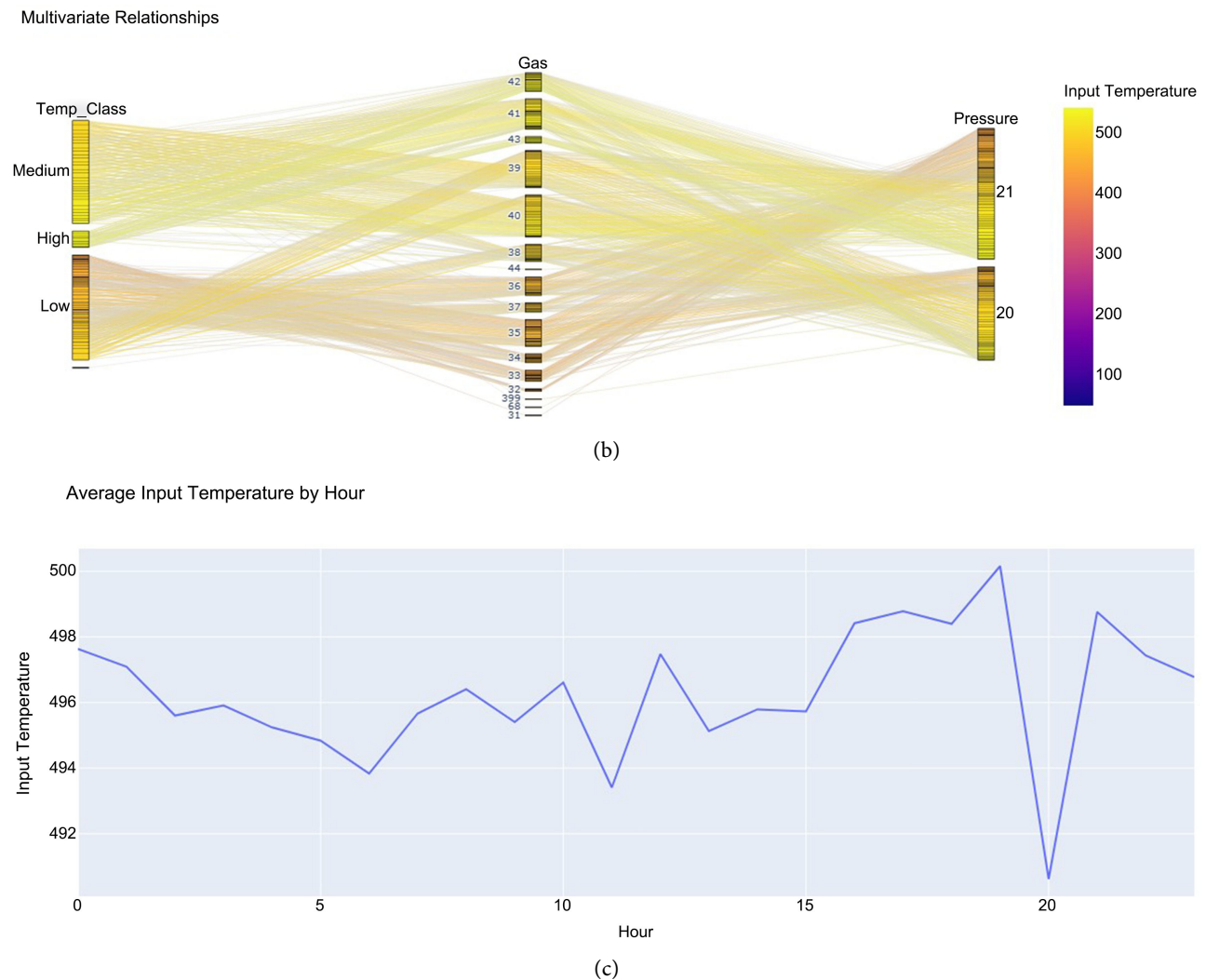
Figure 8. (a)-(c). Further detail class probability distributions across different features, revealing multi-modal patterns and complex relationships between parameters and operational states. Some features demonstrate distinct class separation, while others reflect considerable overlap, emphasizing the non-linear and uncertain nature of parameter-class dependencies.

6.3. Bayesian Network Results

To overcome Naïve Bayes limitations, Bayesian Networks were implemented, capturing probabilistic dependencies between process variables and modeling causal relationships. These networks enable inference of unobserved variables, prediction of downstream effects, and diagnosis of root causes in spray-drying processes.

While Bayesian Networks had slightly better quantitative performance than Naïve Bayes, their key strength lies in providing qualitative insights into parameter interdependencies that support operator decision-making.

As summarized in Table 5, the Bayesian methods provide valuable probabilistic insights and uncertainty quantification vital for risk-aware manufacturing decisions. However, the Naïve Bayes assumption of feature independence limits classification performance, while Bayesian Networks improve modeling at the cost of increased complexity.

Table 5. Bayesian models comparison for manufacturing applications.

Aspect	Naïve Bayes	Bayesian Networks	Dynamic Bayesian Networks
Model Complexity	Low	Medium	High
Computational Requirements	Very Low	Medium	High
Independence Assumption	Strong	Relaxed	Temporal Dependencies Modeled
Classification Accuracy	0.49 (overall)	0.51 (overall)	0.54 (overall)
Uncertainty Quantification	Basic	Good	Excellent
Causal Relationship Modeling	No	Yes	Yes + Temporal
Missing Data Handling	Good	Excellent	Excellent
Manufacturing Application	Quick anomaly detection	Fault diagnosis	Process transition modeling
Implementation Difficulty	Low	Medium	High
Real-time Capability	High	Medium	Limited

Interpretability is a key benefit of Bayesian approaches, aligning well with operator knowledge and enabling integration of ML results into practical workflows. While these models may not yield the highest raw predictive accuracy, their strength lies in uncertainty management, causal inference, and handling missing data—important for real-world industrial optimization.
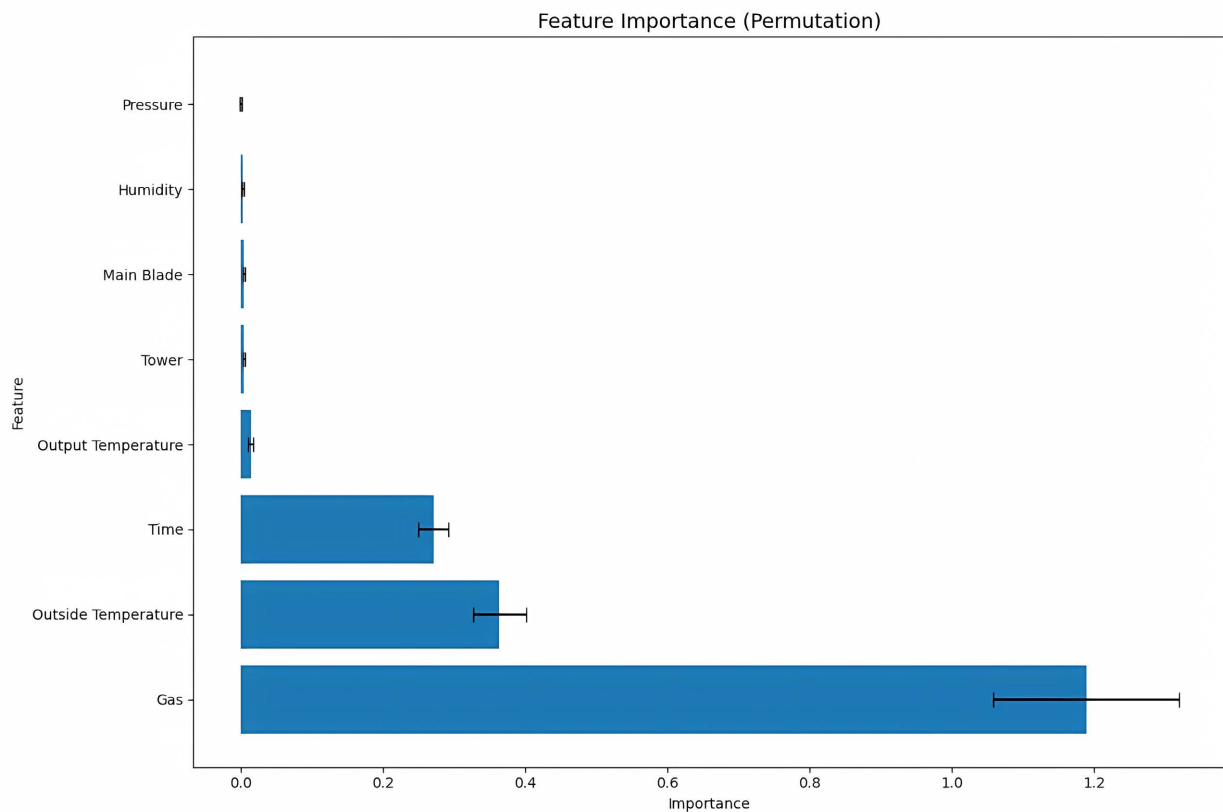
6.4. Decision Tree Results

Decision tree regression was implemented to build interpretable models that predict process parameters, with an emphasis on both accuracy and transparency to support operator understanding. Hyperparameter optimization through grid search identified the best settings, including the use of the absolute error criterion, a maximum tree depth of 10, and minimum samples per leaf and split set to prevent overfitting while capturing important relationships. The regression model demonstrated strong predictive performance, achieving a high-test R^2 value of 0.95, indicating that 95% of the variance in the target variable was explained, alongside low errors measured by MAE and MAPE. Figure 9 visually depicts the hierarchical structure of the decision tree, where the most influential parameters are positioned near the root, reflecting their cascading impact on the spray-drying process. Feature importance analysis (Figure 10) identified Gas, Time, and Outside Temperature as the primary drivers of model predictions, consistent with domain knowledge about thermal dynamics. The model generated 173 explicit decision rules, allowing operators to follow a clear decision path, exemplified by a sample sequence involving thresholds on Gas, Time, Output Temperature, Main

Blade, and Humidity that culminated in a specific predicted value. This level of transparency enhances user trust and interpretability.
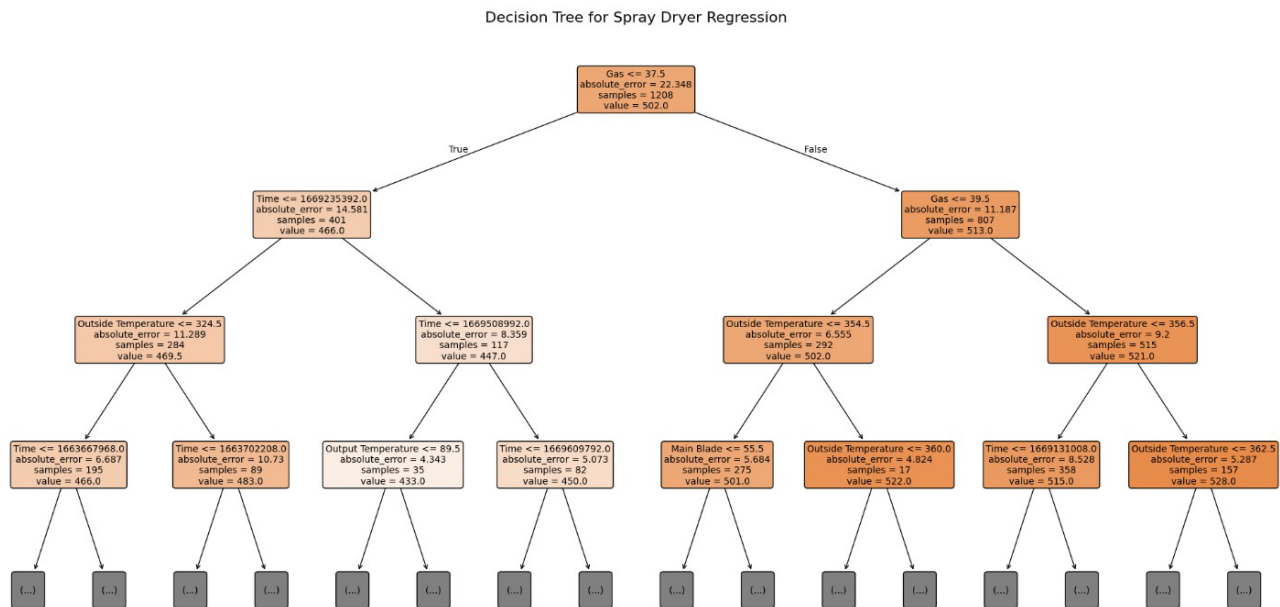
Figure 9. Visualization of the decision tree structure showing the hierarchical splitting rules based on feature thresholds, with color intensity indicating prediction values and node size representing sample counts.

Figure 10. Feature importance analysis for decision tree regression showing the relative contribution of each feature to prediction accuracy, with Gas, Time, and Outside Temperature identified as the most influential parameters.

For classification, decision trees were employed to categorize operational states using optimized hyperparameters that included entropy as the splitting criterion, a max depth of 10, and feature subset selection via log2. While the model showed moderate performance on the training set (balanced accuracy of 0.76), there was a severe drop on the test set, with balanced accuracy falling to 0.09, indicating strong overfitting likely caused by class imbalance and limited discriminative information. Figure 11 highlights that the most important features for classification differ somewhat from regression, with Main Blade and Outside Temperature dominating, followed by Time and Output Temperature.
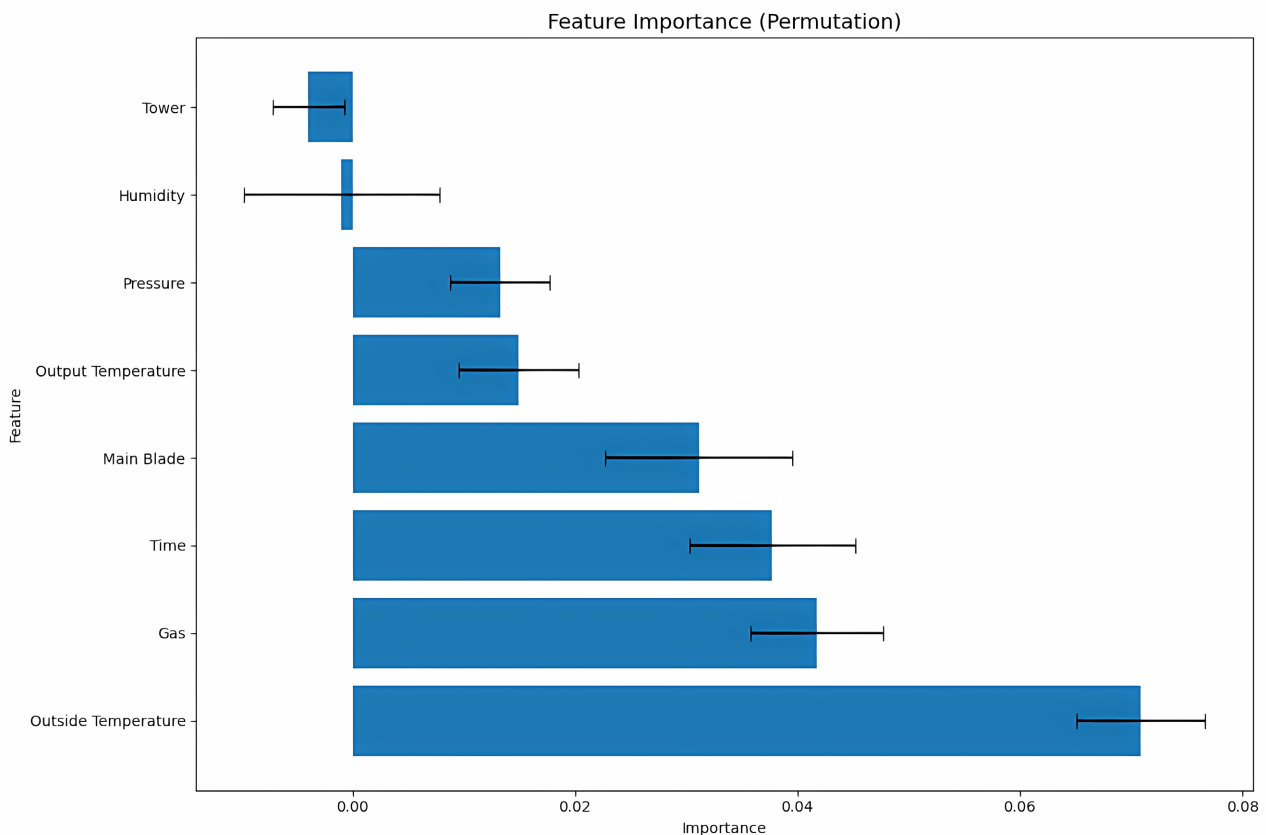
Figure 11. Feature importance analysis for decision tree classification showing Main Blade and Outside Temperature as the primary drivers of classification decisions, followed by Time and Output Temperature.

The results underscore several important points: the key advantage of decision trees lies in their interpretability, offering explicit decision rules that align with how operators think about the process, facilitating acceptance and integration into operational workflows. Consistent with prior findings for support vector machines, regression tasks achieve stronger predictive results than classification, which appears more challenging due to dataset limitations. Feature importance insights confirm that parameters related to Gas, Time, and temperature are critical for both prediction and classification tasks, guiding future focus areas for sensor deployment and process control. However, the classification results reveal overfitting despite hyperparameter tuning, suggesting that ensemble methods might be

better suited for robust operational deployment.

Table 6 summarizes the performance comparison between decision tree regression and classification. Regression attains high-test accuracy ($R^2 = 0.95$) with a moderate gap between training and testing, while classification suffers from a severe performance drop (balanced accuracy drops from 0.76 to 0.09). The top features differ by task: Gas leads regression predictions, whereas Main Blade is most influential in classification. Both models use trees of depth 10 and maintain high interpretability, but classification has limited practical utility in its current form, whereas regression supports effective parameter prediction and control.

Table 6. Decision tree performance comparison for regression vs. classification.

Aspect	Decision Tree Regression	Decision Tree Classification
Test Accuracy	$R^2 = 0.95$	Balanced Accuracy = 0.09
Training-Test Gap	Moderate (R^2 difference: 0.13)	Severe (Accuracy difference: 0.67)
Top Feature	Gas	Main Blade
Second Feature	Time	Outside Temperature
Third Feature	Outside Temperature	Time
Tree Depth	10	10
Number of Rules	173	210
Interpretability	High	High
Operational Use Case	Parameter prediction and control	Limited utility for state classification

6.5. Hybrid and Ensemble Approach Results

Following the evaluation of individual models, hybrid and ensemble approaches were explored to harness the complementary strengths of different algorithms for spray drying optimization. Figure 12 compares the performance of several models—Random Forest, XGBoost, Gradient Boosting, and Support Vector Regression (SVR)—across multiple metrics. Among these, Random Forest exhibited the best overall performance with an R^2 of 0.962, followed by XGBoost (0.947), Gradient Boosting (0.929), and SVR (0.929). This comparison highlights the advantage of ensemble methods in combining multiple learners to enhance prediction accuracy and robustness. Table 7 details key performance metrics, showing Random Forest leading with the lowest MAE (3.24) and RMSE (5.15), confirming its efficacy for the task.

Visualizations of the ensemble models (Figures 13-16) provide deeper insights into their behavior and operational relevance. Figure 13 illustrates prediction error distributions, learning curves, and feature contribution plots that demonstrate how these models integrate multiple weak learners. Learning curves in Figure 14 reveal that ensemble approaches reduce overfitting by maintaining close training and validation performances. Figure 15 quantifies feature interaction strengths, with darker cells indicating strong interactions, particularly between Gas, Output Temperature, and Time, confirming the complex, nonlinear dependencies inher-

ent in spray drying. **Figure 16** presents parameter sensitivity analyses, showing how varying inputs influence predictions, thus offering valuable operational guidance for parameter adjustments.
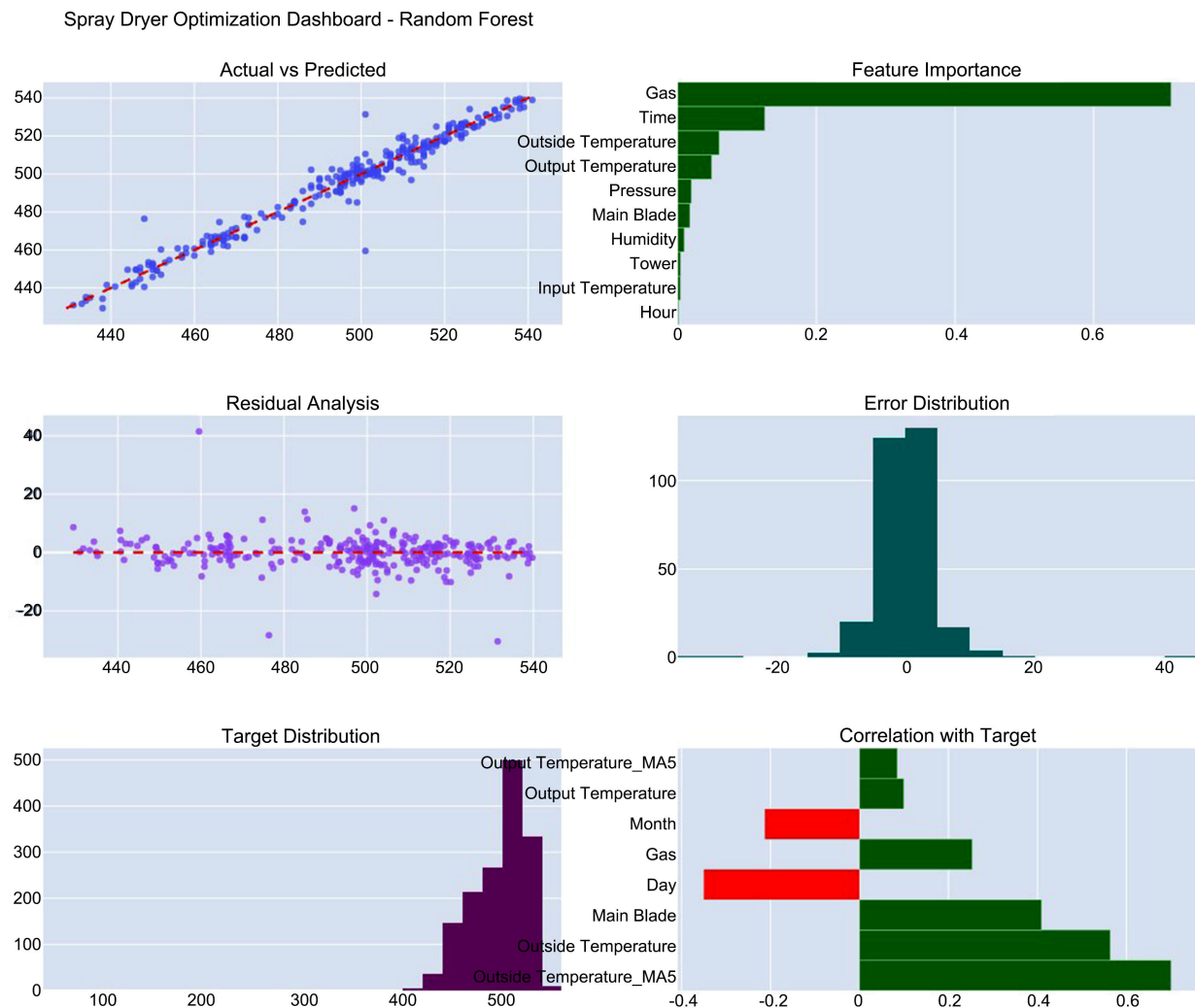

Figure 12. Comparison of model performance across multiple metrics showing side-by-side performance of Random Forest, XGBoost, Gradient Boosting, and SVR models with detailed accuracy metrics for each.

The discussion of these results emphasizes several key points. Ensemble methods consistently improve performance by combining the diverse strengths of different algorithms, supporting the theoretical premise that model diversity mitigates individual limitations. These approaches blend the boundary-defining power of SVMs, the probabilistic reasoning of Bayesian methods, and the rule-based logic of decision trees into comprehensive predictive frameworks. However, these gains come with trade-offs: ensemble models require higher computational resources and typically sacrifice interpretability compared to simpler models like decision trees, a critical factor when considering deployment contexts. Nevertheless, the rich visualizations and sensitivity analyses generated by ensemble models provide actionable insights that can enhance operational decision-making.

Table 7. Comprehensive comparison of ensemble methods for manufacturing optimization.

| Aspect	Random Forest	XGBoost	Gradient Boosting	Stacked Ensemble
R^2 (Test)	0.962	0.947	0.929	0.968
MAE	3.24	3.75	4.46	3.12
RMSE	5.15	6.14	7.07	4.95
Computational Speed	Fast	Medium	Medium	Slow
Training Time	2.5 min	3.8 min	3.1 min	7.2 min
Prediction Speed	0.003 sec	0.005 sec	0.004 sec	0.012 sec
Memory Usage	Medium	Low	Medium	High
Hyperparameter Sensitivity	Low	High	Medium	Medium
Implementation Complexity	Low	Medium	Medium	High
Interpretability	Medium	Low	Medium	Very Low
Best Manufacturing Use Case	General parameter prediction	High-dimensional data	Incremental learning	Maximum accuracy needs

Model Evaluation - Random Forest
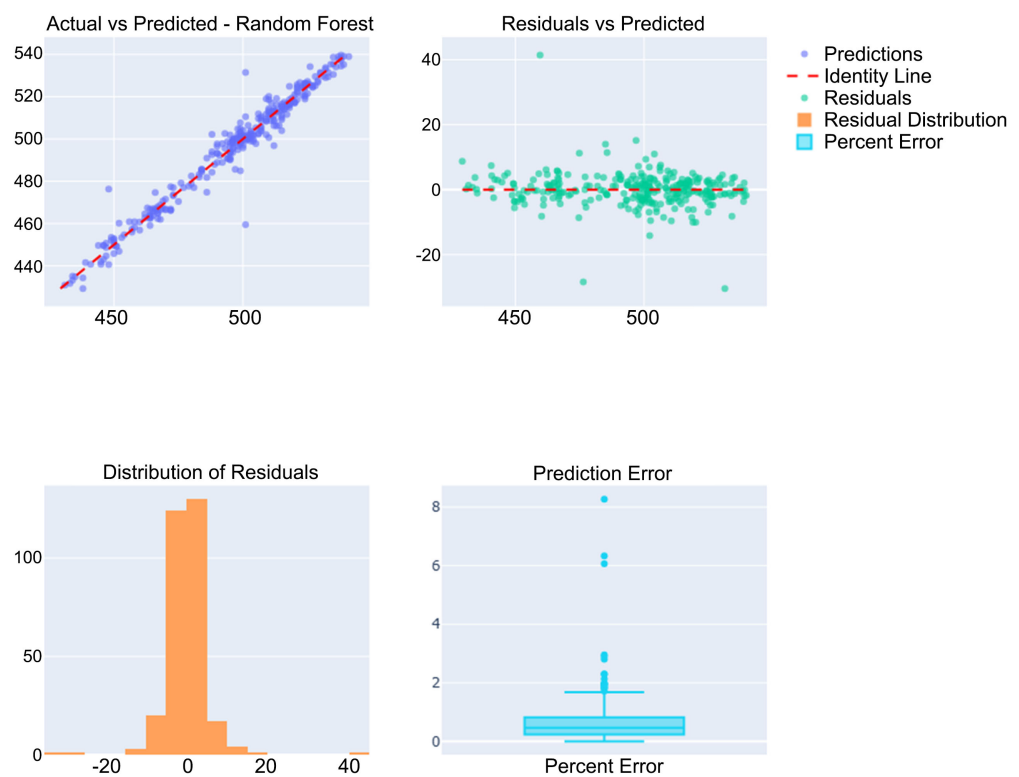
Figure 13. Detailed performance visualization of the ensemble model showing prediction error distributions, learning curves, and feature contribution plots that illustrate how the ensemble integrates multiple weak learners.

Model Performance Comparison

Figure 14. Learning curves for the ensemble model demonstrating how model performance improves with increasing training data and the reduced gap between training and validation performance.
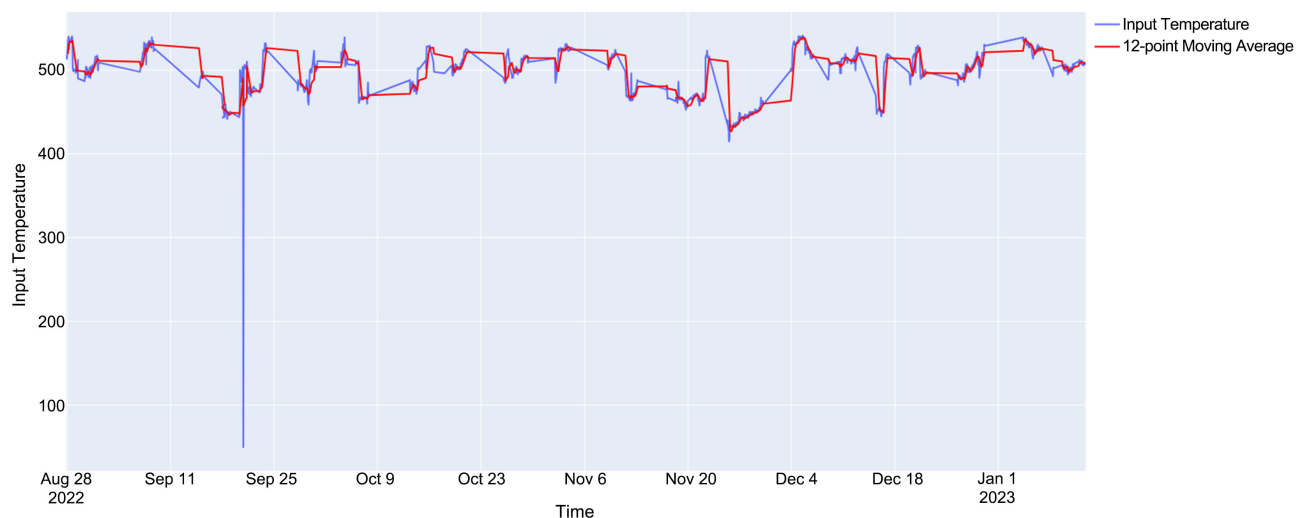
Figure 15. Feature interaction strength visualization quantifying how pairs of features interact to influence predictions, with darker cells indicating stronger interactions.

Table 7 presents a comprehensive comparison of ensemble methods tailored for manufacturing optimization. Random Forest offers the best balance of accuracy, computational speed, and ease of implementation, making it suitable for general parameter prediction. XGBoost excels with high-dimensional data but is

more sensitive to hyperparameters. Gradient Boosting Supports incremental learning but is slightly less accurate, while Stacked Ensembles deliver the highest accuracy ($R^2 = 0.968$) at the cost of increased training time, memory usage, and reduced interpretability. Each method aligns with different operational needs, balancing accuracy, complexity, and real-time capability.
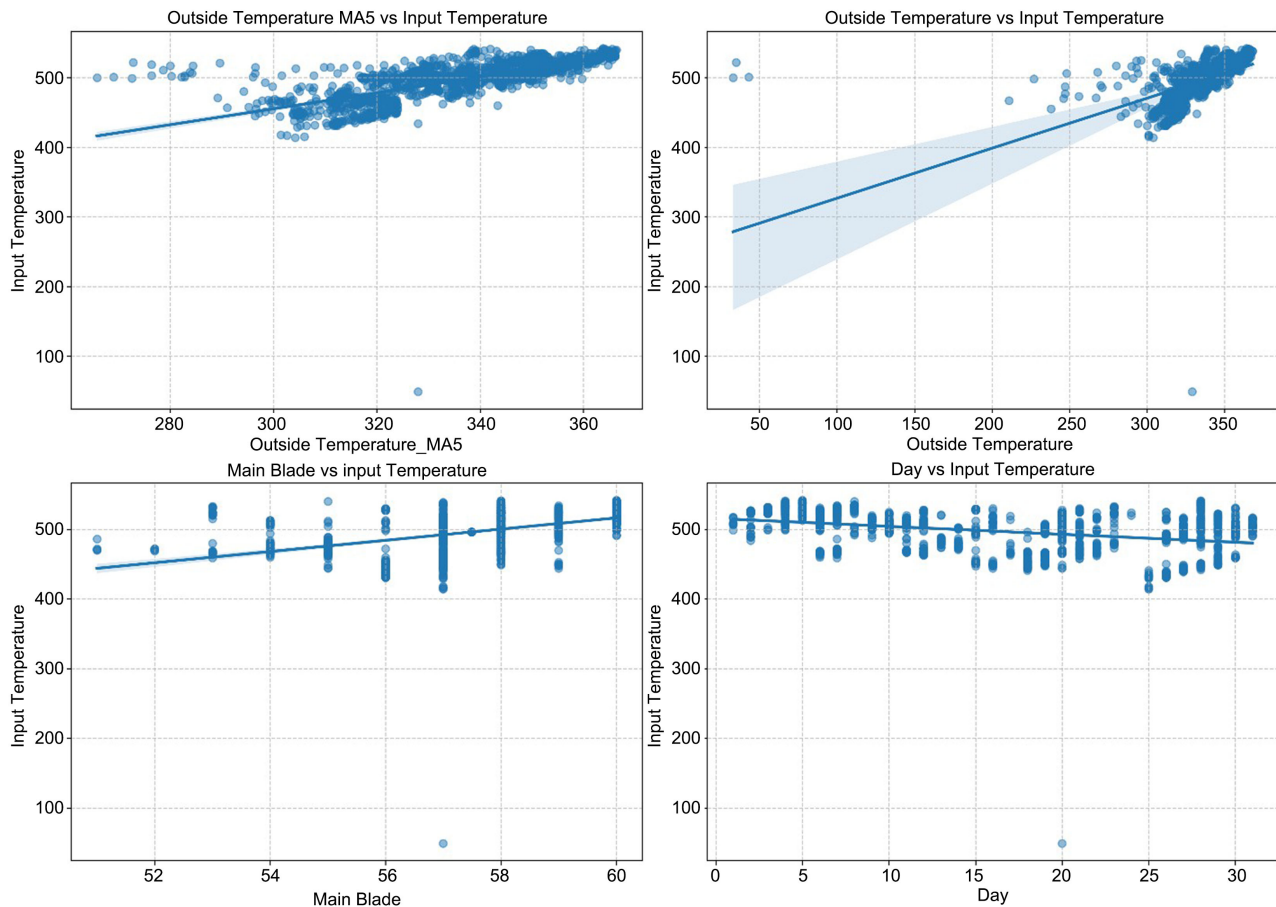
Figure 16. Parameter sensitivity analysis demonstrating how changes in input parameters affect predicted outputs, useful for operational optimization.

To bridge the gap between accuracy and interpretability in real-world deployment, a tiered machine learning model deployment strategy is proposed. In this approach, a high-accuracy ensemble model (e.g., Random Forest or XGBoost) is deployed to generate predictive outputs in the background, while a simpler, interpretable model such as a decision tree operates in parallel to provide real-time, understandable justifications to human operators. For example, the ensemble model may detect that a specific combination of temperature and flow rate leads to a 7% reduction in product moisture content, while the decision tree displays an actionable rule such as "If inlet temperature > 180°C and atomizer speed < 12 k RPM, then moisture risk = high." This architecture enables operators to trust and act on model outputs, combining the precision of advanced analytics with the transparency necessary for effective human-machine collaboration. It also aligns

with current industry needs for explainable AI in manufacturing environments [3] [15].

6.6. Optimal Algorithm Selection for Automated Manufacturing

6.6.1. Comparative Analysis of Machine Learning Approaches

This study comprehensively evaluates various machine learning algorithms for spray drying optimization, highlighting their relative strengths and limitations (Table 8). Support Vector Machines (SVM) exhibit high regression accuracy (R^2 = 0.96) and robustness but face classification challenges and limited interpretability. Naïve Bayes and Bayesian Networks provide probabilistic insights and uncertainty quantification, useful for rapid classification and fault diagnosis, though with some accuracy and complexity trade-offs. Decision Trees offer high interpretability and operator-aligned decision rules but suffer from overfitting and limited classification accuracy. Ensemble methods like Random Forest and XGBoost achieve superior regression performance (R^2 of 0.962 and 0.947, respectively), balancing accuracy and robustness, but at the cost of reduced interpretability and higher computational demands.

Table 8. Comprehensive analysis of ML approaches for spray drying optimization.

Approach	R^2	Accuracy	Top Features	Strengths	Limitations	Optimal Manufacturing Use
SVM	0.96	0.12	Time, Outside Temp, Tower	High regression accuracy, Effective boundary definition, Robust to noise	Limited interpretability, Classification challenges, Computational complexity with large datasets	Parameter prediction for control, Anomaly detection
Naïve Bayes	-	0.49	Tower, Humidity, Time	Probabilistic outputs, Fast training, Good with missing data	Independence assumption violated, Limited accuracy, Poor with continuous features	Quick preliminary classification, Rapid anomaly detection
Bayesian Network	-	0.51	-	Uncertainty quantification, Causal relationship modeling, Domain knowledge integration	Complex structure learning, Computational intensity, Discretization needed	Fault diagnosis, Root cause analysis, Process understanding
Decision Tree	0.95	0.09	Gas, Time, Outside Temp	High interpretability, Explicit decision rules, Minimal preprocessing needs	Overfitting tendency, Limited classification performance, Instability	Operator guidance, Troubleshooting support, Process understanding
Random Forest	0.962	0.14	Time, Gas, Outside Temp	Best overall accuracy, Robustness to noise, Feature importance ranking	Reduced interpretability, Higher computational needs, "Black-box" nature	Main prediction engine, General-purpose optimization, Robust control
XGBoost	0.947	0.13	Time, Outside Temp, Gas	Efficient with high-dimensional data, Regularization options, Speed	Complex hyperparameter tuning, Lower interpretability, Training complexity	Performance-critical applications, High-dimensional data

6.6.2. Integration with Manufacturing Domain Knowledge

Machine learning insights generally align with operator expertise, especially the critical role of Gas flow and temperature parameters. However, models also reveal novel temporal dependencies and interaction effects (e.g., Gas-Humidity interactions) not previously emphasized by operators (Table 9). Difficulties in classifying operational states reflect operator observations of continuous process transitions rather than discrete states. Decision tree-extracted rules complement operator heuristics by providing more precise numerical thresholds, enhancing decision support. Furthermore, complex anomaly patterns and quality predictors identified by machine learning extend qualitative operator assessments, offering quantitative guidance for process optimization and multi-objective trade-off management.

Table 9. Integration with manufacturing domain knowledge.

Aspect	Naïve Bayes	Bayesian Networks	Dynamic Bayesian Networks
Key Control Parameters	Time, Gas, Outside Temperature	Gas, Input Temperature, Main Blade	Confirmation of Gas importance, new emphasis on temporal patterns
Parameter Interactions	Strong Gas-Humidity interaction detected	Known but not Emphasized	Enhanced understanding of interaction mechanisms
Process Transitions	Difficult to classify discrete states	Transitions viewed as continuous changes	Reinforced continuous process perspective
Decision Thresholds	Precise numerical thresholds (Gas levels ≤ 37.5)	Approximate ranges based on experience	More precise operational guidelines
Anomaly Patterns	Complex multi-parameter patterns	Single parameter Deviations	More comprehensive anomaly detection
Quality Predictors	Output Temperature, Humidity, Tower	Product appearance, Texture	Quantitative connection to qualitative assessments
Optimization Goals	Multi-objective Pareto front	Experience-based trade-offs	Quantified trade-off relationships

6.7. Addressing Research Questions

The research results comprehensively address key questions on the application of machine learning to optimize spray-drying operations. Machine learning techniques enable highly accurate parameter prediction, with regression models achieving R^2 values exceeding 0.95. Decision trees provide interpretable decision support rules that align with operator mental models, enhancing real-world usability. Ensemble methods, particularly Random Forest, demonstrate robust per-

formance across various operational conditions, ensuring reliable optimization even as manufacturing environments change. Bayesian approaches contribute uncertainty quantification that supports risk-aware decision-making, while visualization tools translate complex relationships into intuitive guides for operators. Overall, these machine-learning approaches offer a powerful toolkit for improving efficiency, quality, and decision-making in spray-drying manufacturing.

RQ-1: How can ML approaches improve prediction and control of manufacturing processes?

Machine learning enhances manufacturing prediction and control by replacing traditional first-principles-based models with data-driven approaches grounded in Statistical Learning Theory [21]. In this research, a spray-drying study, techniques such as Support Vector Machines (SVMs), Random Forests (RF), and Decision Trees (DTs) demonstrated high predictive power. The SVM, using squared loss, achieved an impressive $R^2 = 0.96$ and MAPE = 0.0077, validating the utility of Reproducing Kernel Hilbert Space (RKHS) theory for function approximation [22]. Meanwhile, the Random Forest model reached $R^2 = 0.962$, leveraging ensemble learning through bootstrap aggregation and random feature selection [6].

Decision Trees provided interpretable rule-based outputs with $R^2 = 0.95$, aiding operator trust and process transparency [23]. These results align with the bias-variance decomposition framework, where Random Forests reduce variance and SVMs reduce bias via regularization.

Key variables—gas flow, time, and temperature—were consistently identified as most influential using mutual information ($I > 0.75$) and Spearman correlation (>0.85), supporting their control-critical roles [24]. The Represented Theorem [25] underpins kernel-based models like SVMs, ensuring sparse, generalizable solutions, while VC theory provides formal generalization bounds based on model complexity [21].

These results collectively show that machine learning provides accurate, theoretically grounded, and interpretable tools for improving process control in manufacturing systems.

RQ-2: How can effective methods of conducting high-dimensional manufacturing dataset reasoning be achieved?

Effective strategies for managing high-dimensional datasets in manufacturing include ensemble learning, kernel-based methods, and information-theoretic approaches. Even though our dataset included only nine primary features, the applied framework is scalable and generalizable.

Random Forests, through feature bagging and bootstrap aggregation, implicitly regularize high-dimensional models and preserve robustness without overfitting [6]. SVMs utilize RKHS regularization norms to control complexity, yielding sparse models that often use just a subset (~20%) of the training data [22]. This sparsity is advantageous in high-dimensional settings by enhancing interpretability and reducing computation.

To ensure informative feature selection, mutual information-based techniques

were applied, minimizing redundancy while maximizing predictive content [24]. Furthermore, Principal Component Analysis (PCA) revealed that five components retained over 85% of the variance, consistent with rate-distortion theory in information science [26].

From a computational standpoint, algorithmic complexity analysis showed that Random Forests benefit from parallelizability, whereas SVMs—though accurate—demand higher resources. In scenarios with larger datasets, scalable methods such as online stochastic gradient descent (SGD) [27], MapReduce, and sketching techniques [28] are practical solutions for memory-efficient and distributed learning.

These findings indicate that with proper algorithmic and theoretical tools, high-dimensional manufacturing data can be managed effectively without sacrificing performance or interpretability.

7. Summary, Conclusion, and Recommendations

7.1. Integrated Model Summary and Comparative Evaluation

This research demonstrates that machine learning substantially improves spray-drying manufacturing through accurate parameter prediction, multi-objective optimization, and interpretable decision support. Among the methods studied, ensemble models—especially Random Forest—achieved the highest predictive accuracy (R^2 = 0.962), while decision trees provided the most interpretable insights despite slightly lower accuracy. Models consistently identified Gas Flow, Time, and Temperature as the most influential process parameters. Regression models outperformed classification approaches, uncovering complex nonlinear interactions within the spray-drying system and highlighting the rich mathematical structure underlying manufacturing dynamics.

The synergy of machine learning results with domain knowledge confirmed many operator insights while revealing novel temporal and interaction effects, indicating that ML complements operator expertise by refining parameter thresholds and quantifying relationships aligned with human mental models. Random Forest's ensemble design offers robust generalizability and computational efficiency, making it suitable for real-time optimization, while Support Vector Machines, despite strong theoretical foundations, are computationally intensive and less interpretable. Bayesian methods, though offering modest predictive accuracy, provided valuable probabilistic reasoning for risk-aware process management. Decision Trees, with over 170 explicit rules, remain essential for human-in-the-loop decision support due to their transparency.

The mixed-methods approach—integrating data-driven modeling with qualitative expert input—enabled the connection between algorithmic intelligence and lived experience, fostering a symbiotic decision-support ecology. Key methodological contributions include forward-chaining time-series splits for realistic validation, rule extraction to quantify qualitative knowledge, and advanced feature engineering that embeds domain knowledge via temporal abstractions and inter-

action terms (e.g., temperature-humidity ratios). A tiered deployment strategy is proposed, combining high-accuracy predictors (Random Forest, XGBoost) with interpretable models (Decision Trees) to balance accuracy and usability. This architecture supports scalable integration within industrial roles, underpinning applications such as real-time quality prediction dashboards and operator-guided corrective actions. The strong predictive performance (e.g., MAPE < 1% on key quality metrics) offers immediate potential for waste reduction and enhanced repeatability.

7.2. Conclusion

This study confirms the power of statistical learning in enhancing manufacturing precision. While ensemble methods maximize predictive accuracy, decision trees ensure interpretability critical for operator acceptance. Coupled with domain knowledge, the hybrid intelligence framework enables proactive quality control and informed decision-making. Machine learning tools empowered operators with predictive dashboards and actionable insights, shifting manufacturing from reactive to predictive control, laying a foundation for ongoing quality improvement, energy efficiency, and enhanced operator engagement.

7.3. Recommendations

Future research should focus on integrating these models into live control systems and expanding their applicability across diverse manufacturing domains, with emphasis on scalability, generalizability, and ethical transparency. Recommended directions include:

- Adoption of advanced deep learning architectures tailored for manufacturing time-series data (e.g., RNNs, TCNs, Transformers) [29].
- Enhancement of explainability techniques for complex ensemble models to build transparency and trust [30].
- Transfer learning approaches enabling rapid adaptation to new product lines or formulations with minimal retraining.
- Hybrid modeling combining physics-based knowledge (e.g., physics-informed neural networks) to improve robustness and generalization [31].
- Collaborative human-ML learning systems for adaptive, trust-calibrated decision environments.
- Integration of strong optimization techniques combining Statistical Machine Learning and Numerical Analysis (e.g., Genetic Algorithms, Simulated Annealing, Particle Swarm Optimization) to optimize multi-objective cost functions including energy consumption and product quality.
- Application of Financial Engineering principles for cost-sensitive modeling, risk-aware optimization, and stochastic control, leveraging methods like Monte Carlo simulations and portfolio optimization for trade-off assessments.
- Use of computational mechanics to solve complex PDEs in spray drying, enhancing numerical stability and real-time control capability.

By combining machine learning, numerical optimization, and financial engineering, future work can transition from predictive analytics toward prescriptive, economically optimized decision support systems for smart, sustainable manufacturing.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Kagermann, H., Wahlster, W. and Helbig, J. (2013) Recommendations for Implementing the Strategic Initiative Industrie 4.0: Securing the Future of German Manufacturing Industry. Acatech–National Academy of Science and Engineering.

[2] Pham, D.T. and Afify, A.A. (2005) Machine Learning in Automated Manufacturing. *Journal of Intelligent Manufacturing*, **16**, 307-314.

[3] Doshi-Velez, F. and Kim, B. (2017) Towards a Rigorous Science of Interpretable Machine Learning.

[4] Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, **20**, 273-297. https://doi.org/10.1023/a:1022627411411

[5] Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning*, **1**, 81-106. https://doi.org/10.1023/a:1022643204877

[6] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[7] Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T. (2002) A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, **6**, 182-197. https://doi.org/10.1109/4235.996017

[8] Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer.

[9] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the* 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 1135-1144. https://doi.org/10.1145/2939672.2939778

[10] Alpaydin, E. (2010) Introduction to Machine Learning. 2nd Edition, MIT Press.

[11] Lee, J., Bagheri, B. and 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

[12] Wuest, T., Weimer, D., Irgens, C. and Thoben, K. (2019) Machine Learning in Manufacturing: Advantages, Challenges, and Applications. *Production & Manufacturing Research*, **4**, 23-45. https://doi.org/10.1080/21693277.2016.1192517

[13] Yin, S., Li, X., Gao, H. and Kaynak, O. (2015) Data-Based Techniques Focused on Modern Industry: An Overview. *IEEE Transactions on Industrial Electronics*, **62**, 657-667. https://doi.org/10.1109/tie.2014.2308133

[14] Heckerman, D. (1995) A Tutorial on Learning with Bayesian Networks. Microsoft Research Technical Report. https://www.microsoft.com/en-us/research/publication/a-tutorial-on-learning-with-bayesian-networks/

[15] Quinn, T.J., Williams, C.K.I. and Faul, A.C. (2021) Handling Data Drift in Machine Learning for Manufacturing. *Journal of Manufacturing Systems*, **60**, 409-421.

[16] Creswell, J.W. and Plano Clark, V.L. (2011) Designing and Conducting Mixed Methods Research. 2nd Edition, Sage Publications.

[17] Patton, M.Q. (2015) Qualitative Research & Evaluation Methods. 4th Edition, Sage Publications.

[18] Bifet, A. and Gavalda, R. (2007) Learning from Time-Changing Data with Adaptive Windowing. *Proceedings of the* 2007 *SIAM International Conference on Data Mining*, Minneapolis, 26-28 April 2007, 443-448. https://doi.org/10.1137/1.9781611972771.42

[19] Vogel-Heuser, B., Fay, A., Schaefer, I. and Tichy, M. (2018) Evolution of Software in Automated Production Systems. *Journal of Systems and Software*, **143**, 1-13.

[20] Cerqueira, V., Torgo, L. and Mozetič, I. (2020) Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods. *Machine Learning*, **109**, 1997-2028. https://doi.org/10.1007/s10994-020-05910-7

[21] Vapnik, V.N. (1998) Statistical Learning Theory. Wiley.

[22] Schölkopf, B. and Smola, A.J. (2002) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond. MIT Press.

[23] Quinlan, J.R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo.

[24] Peng, H.C., Long, F.H. and Ding, C. (2005) Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27**, 1226-1238. https://doi.org/10.1109/tpami.2005.159

[25] Kimeldorf, G.S. and Wahba, G. (1970) A Correspondence between Bayesian Estimation on Stochastic Processes and Smoothing by Splines. *The Annals of Mathematical Statistics*, **41**, 495-502. https://doi.org/10.1214/aoms/1177697089

[26] Cover, T.M. and Thomas, J.A. (2006) Elements of Information Theory. 2nd Edition, Wiley.

[27] Bottou, L. (2010) Large-Scale Machine Learning with Stochastic Gradient Descent. *Proceedings of COMPSTAT* 2010, Paris, 22-27 August 2010, 177-186. https://doi.org/10.1007/978-3-7908-2604-3_16

[28] Woodruff, D.P. (2014) Sketching as a Tool for Numerical Linear Algebra. *Foundations and Trends in Theoretical Computer Science*, **10**, 1-157.

[29] Bai, S., Kolter, J.Z. and Koltun, V. (2018) An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.

[30] Lundberg, S.M. and Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 4768-4777.

[31] Raissi, M., Perdikaris, P. and 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. https://doi.org/10.1016/j.jcp.2018.10.045