Separation and Purification Technology 393 (2026) 137337

Contents lists available at ScienceDirect

Separation and Purification Technology

journal homepage: www.elsevier.com/locate/seppur

CAD to characterization: Machine learning on experimental data for additively manufactured helical distillation columns

Vignesh Jayavelu [a,*], Christoph Klahn [a,b]

[a] Institute for Micro Process Engineering (IMVT), Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Germany
[b] Institute for Mechanical Process Engineering (MVM), Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany

ARTICLE INFO

Communicated by Cintia Marangoni

Keywords:
Machine learning
Helical distillation column
Additive manufacturing
Number of stages
Process intensification

ABSTRACT

Traditional workflows in chemical engineering, from design to characterization, have served the field well. Emerging technologies, such as machine learning, now offer opportunities for faster design iterations and better integration of the various development stages. This study demonstrates how machine learning (ML) can reduce this gap by directly integrating experimental data into predictive models. Focusing on a Additive Manufactured helical micro-distillation column design for separating liquid mixtures, several ML models were trained to predict the number of theoretical stages based on geometric parameters and operating conditions. 11 variants were designed, built, and tested to collect 197 experimental data sets to assess the feasibility of five predictive ML models from experimental data. Among the algorithms tested, the Gradient Boosting Regressor achieved the best performance with a coefficient of determination $R^2 = 0.7934$. The work highlights the behavior of the model across different regimes, identifies key sources of error, and proposes a hybrid experimental–ML workflow for rapid screening of distillation designs. This approach accelerates process development and reduces the need for extensive experimentation, especially in time-consuming tasks such as distillation.

1. Introduction

The design and optimization of distillation columns remain central to efficient chemical separation processes, particularly in compact, decentralized plants where space and energy efficiency are critical. One such emerging application is offshore Power-to-X (PtX) container plants, where processes such as methanol distillation must be implemented within constrained volumes such as standard shipping containers [1]. Among innovative alternatives to conventional packed beds, helical distillation (HeliDist) columns—especially those fabricated via additive manufacturing (AM) have shown potential in improving mass transfer performance while reducing the footprint of the system [2–4]. Additive manufacturing enables more complex shapes that exceed the capabilities of numerical characterization. However, experimental characterization of these structures is time-consuming and resource-intensive due to their complex internal geometries. Complex geometries typically have non-standard flow regimes, and each geometry has to be tested irrespective of its similarity in shape and size.

Fig. 1 illustrates the time consumed across the full workflow from design to final experimentation, recorded by Grinschek et al. [4]. A parametric CAD model, Design for Additive Manufacturing and hybrid unmanned additive manufacturing reduced the lead time of a new variant below 2 days. The experimental characterization is however significantly slower. This study proposes a machine learning (ML) based approach to accelerate characterization by predicting key performance indicators (KPIs) directly from design parameters, thereby reducing reliance on exhaustive experimental screening during design space exploration.

In conventional systems, as described in Fig. 2 KPIs such as the number of theoretical stages (NTS) ⑤ can be estimated through empirical correlations or simulations. However, in AM-HeliDist columns, such correlations fail due to complex geometries, introducing non-linear vapor–liquid interactions and non-analytical pressure drops. Compared to first-principles simulations which require detailed assumptions, are computationally expensive, and often fail to capture manufacturing and post-processing imperfections given in ② and ③. By training directly on experimental data, ML models inherently capture the cumulative effect of real-world disturbances given in ④, including heat loss, calibration errors, and geometric irregularities, without requiring explicit modeling. While the objective of this data-driven approach is to learn the cumulative influence of such real-world disturbances directly from experimental data, the extent to which these effects can be captured is inherently constrained by the coverage and resolution of the available dataset. Fig. 2 illustrates how the influence of these effects can be predicted from ① to ⑤ using ML. This makes ML especially powerful in systems involving AM structures with high fabrication variability.

* Corresponding author.
E-mail address: vignesh.jayavelu@kit.edu (V. Jayavelu).

https://doi.org/10.1016/j.seppur.2026.137337
Received 3 December 2025; Received in revised form 3 February 2026; Accepted 18 February 2026
Available online 20 February 2026

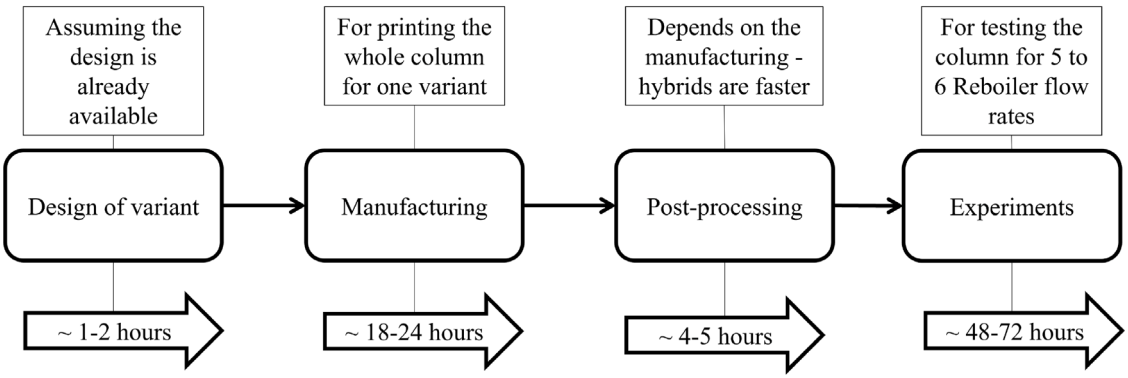

Fig. 1. Timeline for the manufacturing and testing process of AM-based distillation columns [4].

Table 1
Literature review on Machine-learning applications in distillation.

Reference	Process/System	Predicted parameter (s)	ML model (s)
Pullanikkattil et al. (2025) [5]	Binary distillation column (ethane–ethylene)	Top and bottom product composition	XGBoost; SHAP-based interpretability
Negri et al. (2025) [8]	Vacuum distillation column (fouling-prone industrial unit)	Column pressure differential (ΔP)	Gaussian Process Regression (GPR)
Adebayo et al. (2025) [6]	Debutanizer distillation column	Differential pressure trajectory; flooding onset	TimeGAN; supervised ML classifiers
Eldi et al. (2025) [7]	Industrial distillation column	Impurity concentration in distillate	Multiple regressors; LSTM
Elsheikh et al. (2023) [9]	Water distillation tower	Column temperature; water production rate	RVFL neural network
Almahfoodh et al. (2023) [10]	Membrane distillation module	Flux; separation performance indicators	Deep neural network (DNN)
Park et al. (2022) [11]	Industrial distillation process	Product-stage temperature; energy demand	Supervised ML model
Kwon et al. (2021) [12]	Industrial distillation column	Product-stage temperature	Artificial neural network (ANN)

Machine learning has been increasingly applied to distillation processes for monitoring, prediction, and optimization of operational variables. Recent studies have demonstrated the use of data-driven models for predicting product composition in binary distillation columns [5], detecting flooding and pressure-drop evolution [6], estimating impurity levels in industrial columns [7], and assessing fouling-related pressure differentials in vacuum distillation units [8]. A summary of representative machine-learning applications in distillation is provided in Table 1.

As evident from Table 1, existing machine-learning studies in distillation predominantly focus on predicting operational, quality, or maintenance-related parameters such as product composition, temperature, pressure drop, or fouling indicators and not on the internal column geometry—particularly for additively manufactured structures

To address this challenge, several supervised learning algorithms were evaluated for predicting key performance indicators (KPIs) in AM-HeliDist columns. Tree-based ensemble methods such as Random Forest (RF) and Gradient Boosting (GB) have been applied in chemical engineering due to their robustness against noisy experimental data and their ability to model nonlinear input–output relations [13,14]. Extreme Gradient Boosting (XGB), an optimized variant of gradient boosting, further enhances prediction accuracy by incorporating regularization and handling high-dimensional feature spaces [15].

Artificial Neural Networks (ANNs) have a long history in chemical engineering, particularly for capturing complex nonlinear dynamics in reaction systems and separation processes [16]. More recently, Convolutional Neural Networks (CNNs) have also been explored in process engineering applications to extract spatial dependencies from structured or image-based representations of chemical systems [17]. These models together form a diverse supervised learning toolkit that can accelerate characterization and design optimization of novel AM-based separation equipment.

The models are trained on existing experimental data generated from multiple configurations and benchmarked using standard regression metrics. While ML has been applied in conventional distillation [6,7,9–12,14], this work uniquely focuses on AM HeliDist columns using real experimental data [4], addressing the gap between simulation-driven assumptions and manufacturing-driven realities. In many industrial chemical processes, large volumes of operational data have been collected over years for monitoring and control purposes, often before the advent of machine learning. By applying ML techniques to such historical datasets, researchers effectively replicate the real-world scenario of leveraging legacy data to enable predictive modeling and optimization. This approach reflects a pragmatic integration of data-driven tools into existing process frameworks, allowing insights into complex phenomena without requiring extensive new experimentation.

1.1. Machine learning in process engineering

Machine learning (ML) is increasingly used in process systems engineering for tasks such as property prediction, fault detection, process control, and optimization [6,7,10–12,14]. The heterogeneity and volume of data encountered in chemical engineering ranging from simulation outputs to experimental measurements make ML an attractive complement to traditional modeling approaches [18]. Early methods focused on linear regression and support vector machines, while more recent work has shifted toward ensemble techniques and neural networks to capture complex, nonlinear system behavior.

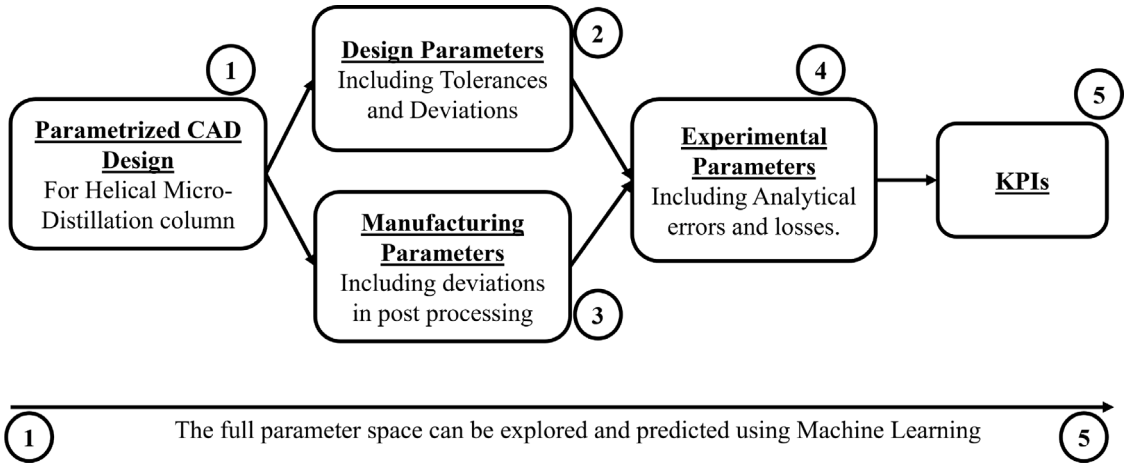

Fig. 2. Flowsheet describing the types of parameters and their influence on KPIs in characterizing an AM HeliDist column.

Schweidtmann and Mitsos [18] emphasize the growing use of ML for surrogate modeling, optimization, and control, particularly when first-principles models are too slow or impractical. However, they also highlight challenges such as model generalization and interpretability, both are critical for safety and deployment in industrial contexts. In response, hybrid modeling approaches have gained attention. For instance, Bikmukhametov and Jäschke [19] propose integrating first-principles knowledge into ML pipelines, improving both transparency and reliability in applications like virtual flow metering. Similarly, Venkatasubramanian [14] explores the use of deep learning and reinforcement learning for process automation while cautioning that data quality and validation remain pressing concerns. These developments suggest a paradigm shift toward more adaptive and intelligent workflows in chemical engineering.

Recent work by Rentschler et al. [20] has shown how recurrent neural networks can accelerate the design of Power-to-X process chains under transient conditions. This highlights the broader role of ML in supporting dynamic process optimization and complements the present study by extending data-driven methods to energy transition applications. Similarly, Kaya et al. [21] demonstrated a machine learning-assisted design automation workflow for non-uniform flow distributors, underscoring how ML can enable rapid design iteration in complex fluidic systems. Their approach parallels the objectives of this study by integrating ML with additive manufacturing to navigate challenging design spaces.

1.2. ML applications in distillation columns

Recent studies have demonstrated the applicability of machine learning (ML) in conventional distillation systems with encouraging results. Elsheikh et al. [9] employed a Random Vector Functional Link (RVFL) neural network optimized with metaheuristics to predict column temperature and water yield in a distillation tower. In their approach, process operating conditions such as feed composition and column specifications were used as inputs, while the RVFL network provided accurate predictions of the key thermodynamic outputs.

Almahfoodh et al. [10] extended ML applications to membrane distillation, where they developed deep neural networks (DNNs) to support module design. Their framework utilized geometric and operating parameters (e.g., membrane characteristics, feed flow rates, and operating temperature) as inputs, with module performance metrics such as flux and salt rejection predicted as the outputs. This highlighted ML's capability to accelerate iterative design cycles. In dynamic column operation, Kwon et al. [12] applied Long Short-Term Memory (LSTM) networks for real-time prediction of distillation column temperature. By feeding in time-series operational variables such as pressure, reflux ratio, and flow rates, the model successfully predicted the temperature trajectory of the production stage with high accuracy (R^2 = 0.924). This demonstrated the potential of recurrent neural networks for temporal forecasting in distillation control. At the industrial scale, Park et al. [11] proposed a three-phase ML-based framework consisting of learning, validation, and improvement steps for energy optimization in mixed butane distillation columns. Their predictive model incorporated column operating conditions and process data as inputs and generated two main outputs: the predicted temperature at the product stage, and the required steam flow rate to maintain desired purity. This enabled real-time operational guidance for plant operators.

Further advancements in impurity control were reported by Eldi et al. [7], who developed an ensemble of ML and deep learning models trained on industrial process datasets. After preprocessing through feature selection, noise reduction, and online bias learning, the models predicted impurity levels in the distillate as the output. The best-performing model, an LSTM with convolution smoothing, showed an almost perfect correlation (R^2 = 0.9998) between measured and predicted values. Safety and operability have also been targeted. Adebayo et al. [6] used a combination of Generative Adversarial Networks (GANs) and supervised ML to address flooding detection in packed distillation columns. Synthetic flooding data were generated using GANs to overcome scarcity, and supervised classifiers were trained on sensor time-series data, particularly pressure drop sequences. The models predicted the onset of flooding conditions as outputs, enabling detection up to 19–60 min before flooding occurred.

Together, these studies illustrate the breadth of ML applications, ranging from temperature and impurity prediction to operational safety. However, most works have been limited to individual conventional geometries. Efforts to apply ML to additive manufacturing (AM) enabled structured geometries remain rare. In such systems, the inputs must capture novel geometric descriptors (e.g., helix parameters, lattice density), while the outputs should directly represent process key performance indicators (KPIs). Comparative benchmarking across multiple ML algorithms remains scarce, motivating the present work on AM HeliDist columns.

1.3. Need for this study

Existing machine-learning applications in distillation predominantly target operational variables or product quality metrics in conventional column geometries, while performance prediction for AM HeliDist columns remains challenging due to their geometry-dependent flow behavior, lack of analytical correlations, and strong sensitivity to manufacturing-induced variability

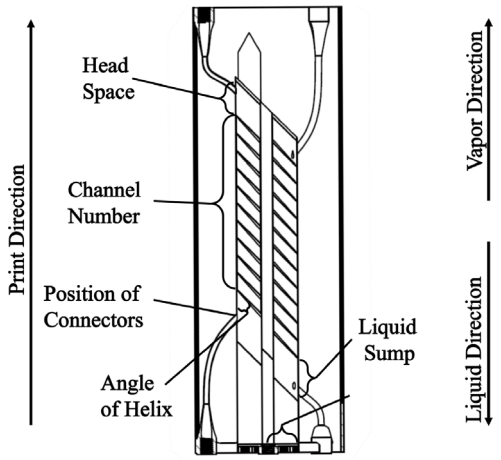

Fig. 3. Representative design features of the AM-based HeliDist distillation column used in this study.

Despite the clear benefits of ML in process engineering, there is currently no focused study that benchmarks ML models for KPI prediction in distillation columns, more particularly for additive manufactured units. These systems pose unique challenges: complex flow paths, geometry-dependent pressure drops, and lack of analytical models make them ill-suited for classical simulation or regression techniques.

This study fills that gap by systematically evaluating the performance of five ML models, including Random Forest, Gradient Boosting, XGBoost, and Artificial Neural Networks, on experimental data from AM-based distillation columns. By doing so, it aims to provide a foundation for ML-assisted design workflows that are better suited for novel and geometrically complex separation systems.

Accordingly, this paper is classified as a data-driven performance prediction study for additively manufactured distillation columns based on already available experimental data (see Fig. 3).

2. Materials and methods

2.1. Design and fabrication of HeliDist columns

The distillation units from Grinschek et al. [2] were designed with a helical flow path to enhance phase interaction within a compact geometry. Parametric 3D models were constructed using Autodesk Inventor®, allowing rapid modification of key design parameters such as channel height and channel number. To ensure compatibility with laser-based powder bed fusion of metal (PBF-LB/M) [22], a 45° tilt of the helix axis was implemented to eliminate the need for internal supports during printing.

The design parameters as given in Fig. 3 such as Head space, Position of connectors/outlets, Angle of helix, Liquid sump, and Column diameter were kept constant throughout the experimental dataset. Only the Channel number (which is the number of Helix turns in the column) and the Channel height (also known as Helix Pitch) were varied as the dominant design parameters. For a helical structure, the hydraulic diameter is an important parameter. This was considered as a factor of Channel height, since the Column diameter was kept constant.

All columns were printed in 316L stainless steel using a Realizer 125 system (DMG Mori, Germany) equipped with a 400 W fiber laser. Wall thicknesses were minimized (approx. 250 µm) to reduce material usage and build time [23]. Post-processing steps included depowdering, sealing surface milling, thread recutting, and leak testing. Two structural variants were prepared: a monolithic design and a modular version with interconnects for cascaded operation. Pneumatic fittings (M5 threads) and machined mounting bases enabled secure installation and thermal insulation [3].

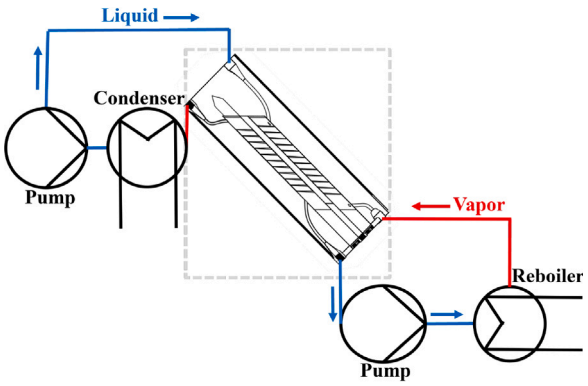

Fig. 4. Experimental process flowsheet for testing the AM distillation columns.

2.2. Experimental setup and operation

Distillation performance was tested under total reflux using an equimolar cyclohexane–heptane mixture in the setup depicted in Fig. 4. The column was mounted at a 45° incline, and heat was supplied via a microstructured evaporator with 35 channels (0.5 × 0.5 × 60 mm^3). A micro annular gear pump (HNP Microsystems, Germany) regulated the reboiler feed. Thermal losses and premature condensation were minimized using PTFE tubing and heat-traced lines.

The vapor stream was condensed using a spiral glass condenser (240 cm^2 surface area), and the liquid was recycled using a second pump. Flow rates and densities were recorded using Coriolis sensors (Bronkhorst, Netherlands), while temperatures were monitored via PT100 sensors. The entire system was controlled by a process automation platform (Hitec-Zang, Germany).

2.3. Key performance indicators (KPIs)

Two KPIs are considered: the Number of Theoretical Stages (NTS) and the Height Equivalent to a Theoretical Plate (HETP). HETP can be calculated based on total column height or unwound helix length. However, since geometric influences are already encoded in the input features, using HETP directly as a target adds redundancy. Therefore, the model focuses on predicting NTS, which is derived from experimental mole fractions of the more volatile cyclohexane in distillate (x_T) and bottoms (x_B) streams using Fenske Equation [24]:

$$\mathrm{NTS} = \frac{\log\left(\frac{x_T}{1-x_T} \cdot \frac{1-x_B}{x_B}\right)}{\log(\alpha)} \quad (1)$$

where:

x_T: Mole fraction of cyclohexane in the top /distillate

x_B: Mole fraction of cyclohexane in the bottoms/residue

α: Relative volatility between cyclohexane and heptane

2.4. Dataset description

The experimental dataset used in this study originates from a complete experimental campaign conducted by a fellow PhD researcher over a period of approximately four years. These experiments were performed prior to the present work and were not designed or conducted with machine-learning applications in mind. In this context, the dataset represents previously generated experimental data within the research group, which is re-analyzed here using data-driven modeling techniques.

After data curation, a total of 197 valid experimental data points comprising 11 variables were retained. Failed experimental runs, incomplete measurements, and statistical outliers were systematically

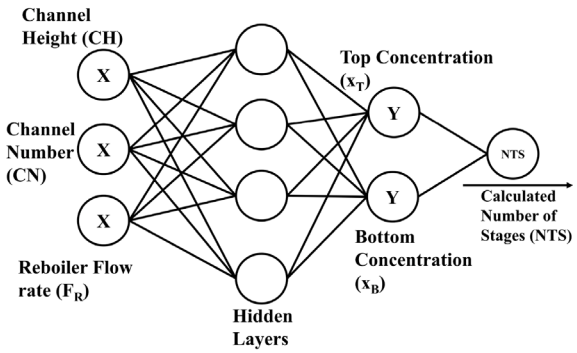

Fig. 5. Flowsheet showing the role of input parameters and their influence on KPIs in AM HeliDist distillation columns.

removed. The dataset includes variations in key design and operating parameters, namely channel height, number of channels, and reboiler flow rate.

Prior to model development, all input features were normalized using Min–Max scaling. The dataset was then randomly divided into training and testing subsets with an 80:20 split. The training set contained data from all 11 column geometries to ensure representative coverage of the design space. The machine-learning models were trained exclusively on the experimentally measured top and bottom concentrations (x_T and x_B). The Number of Theoretical Stages (NTS) was not predicted directly, but was calculated in a post-processing step using the Fenske equation based on the predicted concentration values.

2.5. Machine learning models

Supervised learning models are trained on labeled datasets, where input variables (e.g., flow rate, geometry) are mapped to known output responses (e.g., purity). These models learn functional relationships between variables to make predictions on unseen data [15,25].

In this study, five supervised machine learning (ML) algorithms were benchmarked to predict key process outcomes from experimental and design variables. Random Forest (RF) is an ensemble method based on decision trees that uses bootstrap aggregation, making it robust to noise and effective in modeling non-linearities in process systems [26]. Gradient Boosting (GB) is a sequential tree-based model that incrementally minimizes residual errors and has been successfully applied in yield prediction and fault detection [9]. Extreme Gradient Boosting (XGBoost) is an optimized and regularized version of GB, known for its superior accuracy and computational efficiency in large-scale process datasets [19]. Artificial Neural Networks (ANN) are feedforward models capable of approximating complex input–output relationships in chemical reactors, distillation systems, and other unit operations [16]. Convolutional Neural Networks (CNN), traditionally used for image data, were experimentally explored to capture spatial patterns in structured input features such as channel layout or gyroid surface descriptors [27].

All models were implemented using standard Python machine learning libraries such as scikit-learn, xgboost, and TensorFlow/Keras. Model accuracy was evaluated using the coefficient of determination (R^2), which indicates how much of the variation in the target variable is explained by the model. Hyperparameter tuning was carried out using grid search and random search. For tree-based models, the main parameters adjusted were max_depth, n_estimators, and learning_rate. For neural networks, the key settings optimized included the number of layers, number of neurons, batch size, activation functions, and learning rate, supported by Keras Tuner. Model robustness was ensured through a 5-fold cross-validation strategy.

Table 2
Comparison of performance metrics of machine learning models sorted by decreasing R^2.

Model	R^2	RMSE	MAE	MSE
GB	0.7955	0.3440	0.2580	0.1183
RF	0.7171	0.4048	0.1281	0.1837
XGB	0.6941	0.4186	0.2997	0.1753
ANN	0.6214	0.4661	0.3330	0.2173
CNN	0.5350	0.5193	0.3301	0.2697

3. Results and discussions

3.1. Interpretation of results

Results from this Supervised Machine learning-Regression approach for all the models are summarized in Table 2. Overall, tree-based ensemble methods consistently outperformed the neural-network-based models. This outcome reflects both the characteristics of the dataset and the inherent strengths and limitations of the respective algorithms.

3.1.1. Neural networks

The Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) exhibited the weakest performance. This can be attributed to several factors. First, the dataset size was relatively small, which limited the ability of these high-capacity models to learn robust representations without overfitting. Neural networks typically require thousands of diverse samples to achieve stable convergence and generalization, particularly when input features do not exhibit clear spatial or temporal dependencies.

For CNNs specifically, the poor performance is expected because the tabular, feature-based input structure does not exploit the spatial feature extraction capability for which CNNs are designed. These models could become more competitive if the dataset were expanded and complemented with spatial descriptors of the internal structures (e.g., 3D CAD models, voxelized lattice geometries or CFD-based fields), enabling the network to extract physically meaningful spatial correlations.

3.1.2. Tree-based models

The Tree-based models Gradient Boosting (GB), Random Forest (RF), and Extreme Gradient Boosting (XGB) achieved significantly better accuracy. Their ability to capture complex, nonlinear interactions between geometric and operational parameters without the need for extensive feature scaling or preprocessing makes them well-suited for the current dataset.

Gradient Boosting was selected as the primary model because it achieved the highest R^2 and the lowest RMSE and MSE, indicating superior global predictive performance and fewer large deviations. While the Random Forest model yielded a slightly lower MAE, MAE reflects average absolute deviation and does not penalize occasional large errors as strongly as RMSE. For design-oriented prediction of distillation performance, limiting large mispredictions is critical, as these can lead to incorrect conclusions during design comparison and optimization. Therefore, the model selection prioritized overall variance explanation and robustness against large errors, for which Gradient Boosting performed best.

Fig. 6 compares the predicted and experimental NTS values for the validation dataset. The Gradient Boosting model demonstrates a reasonable predictive capability, capturing the overall trend between experimental and predicted values. However, several data points exhibit noticeable deviations from the ideal parity line, indicating local inaccuracies in model generalization.

These deviations are not random because the dataset used in this study has certain limitations. Some design and operating regimes,

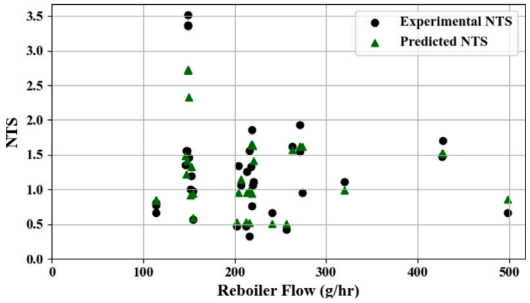

Fig. 6. Experimental vs. predicted results on the validation set using Gradient Boosting.

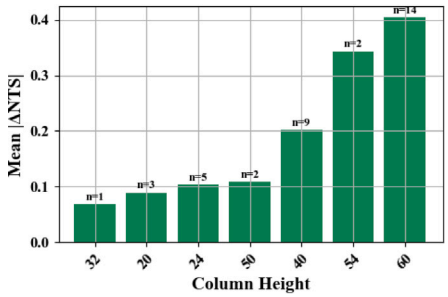

Fig. 7. Model error vs. column height: higher deviations for taller columns (high NTS)

such as high column heights or extreme flow ratios, are underrepresented, which leads to poor generalization when the model encounters these sparsely populated regions during validation. Additionally, several physical effects such as heat losses to the surroundings, dynamic fluctuations during operation, and local maldistribution within larger column heights are not captured in the input features. Their absence reduces the model's ability to accurately describe variations in NTS under complex operating conditions.

As illustrated in Fig. 7, deviations become more pronounced for taller columns associated with higher NTS values. This behavior can be explained by scale-dependent phenomena such as internal maldistribution and increased heat loss effects that are not explicitly parameterized in the current input space.

3.1.3. Impact of the 2+1 output hybrid approach

The multi-output learning strategy, where the model predicts top and bottom concentration, and calculate NTS from the predicted outputs as described in Fig. 5, improved overall performance compared to separate single-output models. This improvement is attributed to the inherent physical coupling between these KPIs. For example, changes in channel height or number of turns influence both top and bottom concentrations, and these relationships collectively inform NTS predictions. By allowing the model to learn these interdependencies in a shared representation, the predictive accuracy and stability of the outputs were enhanced.

This approach essentially embeds domain knowledge into the learning process: instead of treating each KPI as independent, the model leverages correlations within the mass and energy balance of the system, resulting in better generalization, particularly for intermediate operating conditions. In our study, prediction errors were more pronounced in sparsely sampled regimes, such as taller columns associated with higher NTS values. However, given the limited dataset size, we cannot establish a statistically robust correlation between datapoint density and model accuracy

Fig. 8 shows the predicted Number of Theoretical Stages (NTS) surface generated by the Gradient Boosting model across the entire

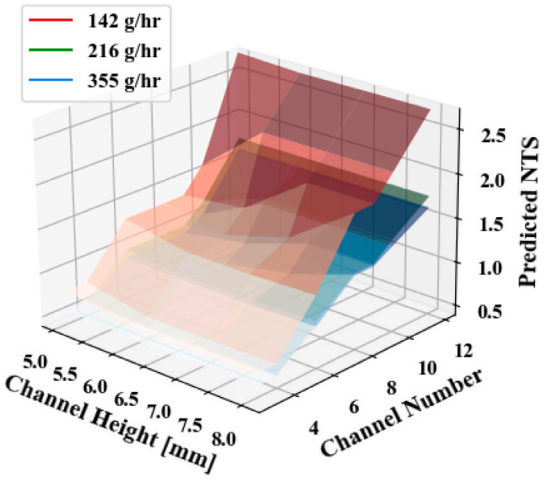

Fig. 8. Predicted Number of Theoretical Stages (NTS) plotted over the geometric design space for three representative reboiler flow rates (142, 216, and 355 g/hr).

geometric and operational design space. The selected reboiler flow rates (142, 216, and 355 g/hr) correspond to representative low, intermediate, and high operating conditions within the experimentally investigated dataset and were chosen to illustrate model behavior across the accessible operating range rather than as optimized setpoints. The three transparent surfaces correspond to different reboiler flow rates and clearly reveal how the model captures nonlinear trends that are difficult to infer from experiments alone. Higher NTS values generally occur at intermediate flow rates and at combinations of channel height and channel number that translate to increased overall column height.

What becomes apparent from the 3D landscape is that the model is able to interpolate well within regions that are densely populated by experimental points, producing smooth and physically consistent gradients. In contrast, flatter regions or abrupt slope changes indicate areas where the model is influenced more strongly by extrapolation than by direct measurements. These transitions highlight the boundaries of the experimental dataset, showing where additional measurements would most effectively reduce uncertainty.

This observation provides the foundation for a hybrid experimental–ML workflow. Rather than relying on exhaustive experimental campaigns, the model's surface can be used to identify informational gaps or regions where the predicted response becomes less stable, or where different flow regimes generate ambiguous behaviors. Targeting these specific areas with new experiments would systematically improve model accuracy while minimizing experimental effort. In this sense, the ML model acts not as a replacement for experiments, but as a guide that prioritizes which experiments are most valuable.

By iteratively updating the model with new high-value data points, the combined ML–experimental approach becomes increasingly efficient. The model continuously refines its understanding of the distillation behavior while experiments focus on the most sensitive, critical, or under-represented regions. This feedback loop forms the core of the hybrid workflow and provides a structured path toward rapid screening and performance optimization of additively manufactured distillation columns.

3.2. Model interpretability and feature influence

To gain insight into the internal behavior of the best-performing Gradient Boosting model and to assess the relative influence of the input variables on the predicted number of theoretical stages (NTS),

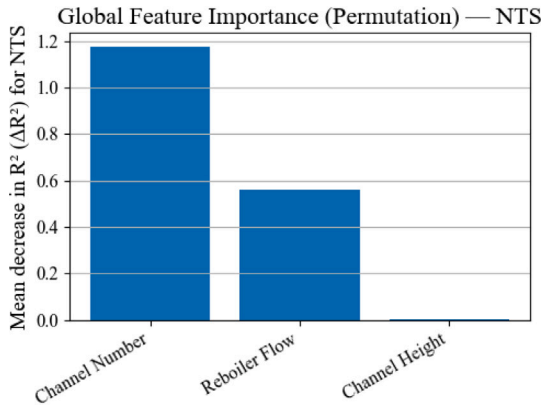

Fig. 9. Permutation-based global feature importance for NTS obtained from the Gradient Boosting model, expressed as the decrease in predictive performance (ΔR^2) upon random permutation of each input feature.

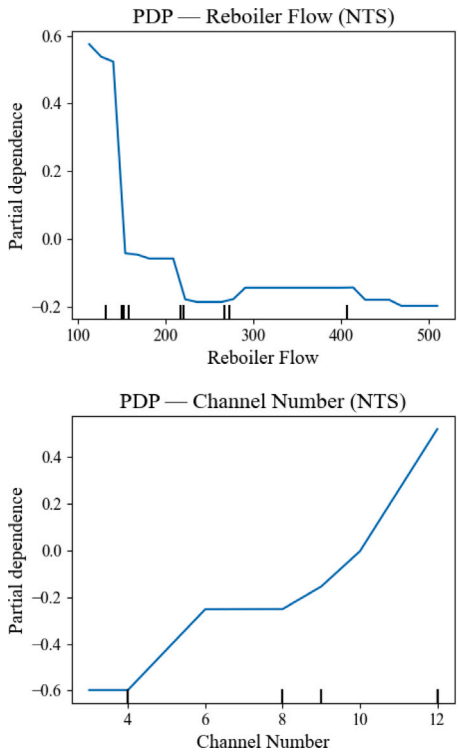

Fig. 10. Partial dependence plots (PDPs) for the Gradient Boosting model showing the effect of (a) Reboiler Flow and (b) Channel Number on the predicted NTS.

permutation-based global feature importance and partial dependence analyses were conducted.

Fig. 9 presents the global feature importance based on permutation analysis, quantified as the decrease in predictive performance (ΔR^2) when individual input features are randomly permuted. The results indicate that the Channel Number is the dominant contributor to NTS prediction, followed by the Reboiler Flow, while the Channel Height exhibits a comparatively minor influence within the investigated design space. This ranking suggests that the overall column extension, represented by the number of helical turns, primarily governs separation performance, whereas operating conditions act as secondary constraints.

To further investigate how individual variables affect the predicted NTS, partial dependence plots (PDPs) are shown in Fig. 10. The PDP for Reboiler Flow (Fig. 10(a)) reveals a pronounced non-linear relationship, with higher predicted NTS values at low flow rates, followed by a sharp decrease and eventual saturation at higher flows. This behavior indicates a transition from a regime dominated by favorable contact conditions to a flow-limited regime in which further increases in vapor flow do not enhance separation efficiency. The PDP for Channel Number (Fig. 10(b)) shows a generally increasing trend, confirming that additional helical turns improve separation performance by extending effective contacting length, albeit with diminishing returns at higher channel numbers. Together, these interpretability results demonstrate that the Gradient Boosting model captures physically plausible trends consistent with established distillation behavior. Importantly, the applied methods provide global and averaged insights into model behavior, which are well suited for the present dataset size and the strongly coupled nature of the geometric design parameters. While advanced explainable artificial intelligence (XAI) techniques such as SHAP-based local attributions can provide detailed insight into individual predictions, their robust application requires larger and more diverse datasets as well as explicit descriptors of relevant physical effects. Given the limited dataset size and the absence of direct indicators for phenomena such as heat losses or maldistribution, local attribution methods may lead to over-interpretation. Within the present scope, permutation-based feature importance and partial dependence analysis provide sufficient and physically consistent insight into the learned model behavior.

4. Conclusion and outlook

This study demonstrates the potential of machine learning (ML) to reduce experimental effort in the design and characterization of additively manufactured helical distillation columns, where rapid iteration and compact design are critical. By modeling intermediate outputs such as the top and bottom product concentrations, rather than relying directly on derived KPIs, like the number of theoretical stages (NTS), the approach improves model robustness and interpretability. Among the tested algorithms, Gradient Boosting delivered the best performance given the dataset's size and complexity.

To further enhance accuracy and generalizability, future work should provide clearer guidelines on the composition of the experimental dataset. Ensuring balanced coverage across the design space nevertheless remains important to reduce extrapolation risks. Moreover, incorporating additional physics-relevant inputs such as pressure drop, wall temperature profiles, or simplified flow descriptors may help capture effects not well represented in the current dataset. Finally, the integration of hybrid models that combine first-principles understanding with data-driven learning holds particular promise for capturing complex behaviors while preserving interpretability. As these models mature, they can serve as intelligent surrogates, accelerating innovation in both conventional and advanced process systems.

Beyond distillation, the workflow developed in this work can serve as a blueprint for other areas of chemical engineering where experimental characterization is slow, costly, or geometrically complex. Additively manufactured heat exchangers, catalytic reactors, membrane systems, or lattice-based absorbers exhibit similarly nonlinear behaviors that are difficult to describe using classical correlations alone. A hybrid experimental–ML loop, in which model uncertainty guides targeted experimentation, can significantly accelerate development cycles for AM-based equipment.

In this context, the integration of explainable AI techniques such as SHapley Additive exPlanations (SHAP) represents a promising extension of the present framework. SHAP-based feature attribution can provide quantitative insight into how individual design and operating parameters influence model predictions, thereby supporting physical interpretability and aiding hypothesis generation. Such interpretability

tools are particularly valuable for additively manufactured systems, where complex geometries and coupled transport phenomena often obscure direct cause–effect relationships. As data-driven surrogates mature and are combined with explainability methods, they may evolve into intelligent design tools enabling predictive screening, rapid prototyping, and informed decision-making in next-generation chemical process development.

Symbols and abbreviations

- CN — Channel number (–)
- CH — Channel height (mm)
- H_c — Column height (mm)
- $\dot{V}$ — Reboiler flow rate (m^3/s)
- $\dot{L}$ — Condenser flow rate (m^3/s)
- T — Temperature (K or °C)
- x_T — Mole fraction of light component in top(distillate)
- x_B — Mole fraction of light component in bottoms(residue)
- α — Relative volatility (–)
- NTS — Number of Theoretical Stages
- $HETP$ — Height Equivalent to a Theoretical Plate (mm)
- ANN — Artificial Neural Network
- CNN — Convolutional Neural Network
- RF — Random Forest
- GB — Gradient Boosting
- XGB — Extreme Gradient Boosting
- ML — Machine Learning
- CFD — Computational Fluid Dynamics
- R^2 — Coefficient of Determination
- ΔNTS — Prediction deviation in NTS

CRediT authorship contribution statement

Vignesh Jayavelu: Writing – original draft, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Christoph Klahn:** Writing – review & editing, Supervision, Conceptualization.

Declaration of Generative AI Use

During the preparation of this work, the authors used ChatGPT to improve readability and proper formatting. After using these tools, the authors reviewed and edited the content as needed and assume full responsibility for the content of this publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Vignesh Jayavelu reports that the financial support was provided by Bundesministerium für Forschung, Technologie und Raumfahrt (BMFTR) in the Hydrogen platform Project H2Mare PtX-Wind under grant 03HY302A. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Dr. Fabian Grinschek for providing the raw experimental data used in this study.

Data availability

The data that has been used is confidential.

References

[1] H2mare: Offshore production of green hydrogen and PtX products, 2026, https://perma.cc/E8KU-RTCE. (Accessed 3 February 2026).
[2] Fabian Grinschek, Jannik Betz, Chen-Mei Chiu, Sören Dübal, Christoph Klahn, Roland Dittmeyer, Additive manufactured helical micro distillation units for modular small-scale plants, Chem. Eng. Process. - Process. Intensif. 208 (2025) 110113.
[3] Fabian Grinschek, Sören Dübal, Christoph Klahn, Roland Dittmeyer, Einfluss des additiven Fertigungsverfahrens auf die Gestalt einer Mikrorektifikationsapparatur, Chem. Ing. Tech. 94 (7) (2022) 958–966.
[4] Fabian Grinschek, Entwicklung additiv gefertigter, modularer, mikrostrukturierter Rektifikationsapparate Dissertation, Karlsruhe Institute of Technology, 2025.
[5] Suhailam Pullanikkattil, Raju Yerolla, Chandra Shekar Besta, Interpretable machine learning model for predicting ethane–ethylene composition in binary distillation process, Therm. Sci. Eng. Prog. 58 (2025) 103236.
[6] Opeoluwa Adebayo, Syed Imtiaz, Salim Ahmed, Machine learning for early detection of distillation column flooding, Chem. Eng. Sci. 312 (2025) 121603.
[7] Gian Pavian Eldi, Ahmad Syauqi, Hankwon Lim, Riezqa Andika, Enhancing distillation column impurity prediction: A novel machine learning and deep learning approach, Ind. Eng. Chem. Res. 64 (26) (2025) 13230–13245.
[8] Francesco Negri, Andrea Galeazzi, Francesco Gallo, Flavio Manenti, Reshaping industrial maintenance with machine learning: Fouling control using optimized Gaussian process regression, Ind. Eng. Chem. Res. 64 (2025) 6633–6654.
[9] Ammar H. Elsheikh, Emad M.S. El-Said, Mohamed Abd Elaziz, Manabu Fujii, Hamed R. El-Tahan, Water distillation tower: Experimental investigation, economic assessment, and performance prediction using optimized machine-learning model, J. Clean. Prod. 388 (2023) 135896.
[10] Sarah Almahfoodh, Adnan Qamar, Sarah Kerdi, Noreddine Ghaffour, Machine learning and computational approaches for designing membrane distillation modules, Sep. Purif. Technol. 325 (2023) 124627.
[11] Hyundo Park, Hyukwon Kwon, Hyungtae Cho, Junghwan Kim, A framework for energy optimization of distillation process using machine learning-based predictive model, Energy Sci. Eng. 10 (6) (2022) 1913–1924.
[12] Hyukwon Kwon, Kwang Cheol Oh, Yeongryeol Choi, Yongchul G. Chung, Junghwan Kim, Development and application of machine learning-based prediction model for distillation column, Int. J. Intell. Syst. 36 (5) (2021) 1970–1997.
[13] Leo H. Chiang, E.L. Russell, Richard D. Braatz, Fault Detection and Diagnosis in Industrial Systems, Springer, 2001.
[14] Venkat Venkatasubramanian, The promise of artificial intelligence in chemical engineering: Is it here, finally? AIChE J. 65 (2) (2019) 466–478.
[15] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning, Springer, 2013.
[16] David M. Himmelblau, Applications of artificial neural networks in chemical engineering, Comput. Chem. Eng. 24 (9–10) (2000) 1127–1139.
[17] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[18] Artur M. Schweidtmann, Erik Esche, Asja Fischer, Marius Kloft, Jens-Uwe Repke, Sebastian Sager, Alexander Mitsos, Machine learning in chemical engineering: A perspective, Chem. Ing. Tech. 93 (12) (2021) 2029–2039.
[19] Timur Bikmukhametov, Johannes Jäschke, Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models, Comput. Chem. Eng. 138 (2020) 106834.
[20] Philipp Rentschler, Stefanie Baranowski, Christoph Klahn, Roland Dittmeyer, Accelerated design of power-to-X process chains for transient operation using recurrent neural networks, Procedia CIRP 128 (2024) 668–673, 34th CIRP Design Conference.
[21] Mertcan Kaya, Julian Ferchow, Mirko Meboldt, Christoph Klahn, Machine learning-assisted design automation workflow of non-uniform flow distributors for multipass crossflow device, Results Eng. 27 (2025) 106931.
[22] Christoph Klahn, Bastian Leutenecker-Twelsiek, Design guidelines, in: Springer Handbook of Additive Manufacturing, Springer, 2023, pp. 177–198.
[23] Fabian Grinschek, Amal Charles, Ahmed Elkaseer, Christoph Klahn, Steffen G. Scholz, Roland Dittmeyer, Gas-tight means zero defects - design considerations for thin-walled fluidic devices with overhangs by laser powder bed fusion, Mater. Des. 223 (2022) 111174.
[24] J.D. Seader, E.J. Henley, D.K. Roper, Separation Process Principles, third ed., Wiley, 2016.
[25] Stuart Russell, Peter Norvig, Artificial Intelligence: A Modern Approach, third ed., Pearson Education, 2010.
[26] Leo Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[27] Shengli Jiang, Victor M. Zavala, Convolutional neural nets in chemical engineering: Foundations, computations, and applications, AIChE J. 67 (9) (2021) e17282.