PharmaQAI: A Machine Learning Framework and Interactive Platform for Process Intelligence and Quality Prediction in Pharmaceutical Tablet Compression

Arsh Chanana , Mithilesh Singh , Mohit Agarwal , Ravindra Pal Singh , Himmat Singh Chawra , Anurag Mishra *

Posted Date: 3 March 2026

doi: 10.20944/preprints202603.0225.v1

Keywords: pharmaceutical tablet manufacturing; machine learning; quality by design; process analytical technology

 doi:10.20944/preprints202603.0225.v1

Article

PharmaQAI: A Machine Learning Framework and Interactive Platform for Process Intelligence and Quality Prediction in Pharmaceutical Tablet Compression

Arsh Chanana, Mithilesh Singh, Mohit Agarwal, Ravindra Pal Singh, Himmat Singh Chawra and Anurag Mishra *

NIMS Institute of Pharmacy, NIMS UNIVERSITY Rajasthan

* Correspondence: anurag.mishra@nimsuniversity.org

Abstract

Predicting pharmaceutical product quality from manufacturing process parameters is a central objective of Quality by Design and Process Analytical Technology frameworks. This study presents a systematic machine learning analysis of 1,005 tablet compression batches characterized by 27 process parameters and six critical quality attributes including drug release, residual solvent, and total impurities. Nine regression and seven classification algorithms were evaluated with randomized hyperparameter optimization and five-fold cross-validation. Tree-based ensemble methods, particularly Extra Trees, consistently outperformed linear approaches across all quality targets. Total impurities achieved the highest predictive accuracy with a test R^2 of 0.8855, driven primarily by formulation-specific categorical identifiers, while drug release targets yielded moderate R^2 values of 0.40 to 0.47, reflecting the inherently complex process-dissolution relationships. Classification of weekend batch production using Logistic Regression yielded an AUC of 0.9215 and cross-validated accuracy of 0.9493, confirming that production schedule characteristics are reliably encoded in process signatures. Feature importance analysis identified tablet fill weight and compaction force as the dominant drivers of dissolution performance. A dedicated Streamlit web application, PharmaQAI, was developed to integrate data exploration, supervised model training, residual diagnostics, and interactive contour-based response surface visualization into a single accessible platform, supporting proactive data-driven decision making in pharmaceutical manufacturing without requiring programming expertise.

Keywords: pharmaceutical tablet manufacturing; machine learning; quality by design; process analytical technology

1. Introduction

Pharmaceutical tablet manufacturing is one of the most complex and tightly regulated production processes in the healthcare industry, requiring the simultaneous control of numerous interrelated process parameters to consistently deliver products that meet stringent quality and safety specifications [1–6]. Tablet compression, in particular, involves a series of mechanically interdependent operations where variables such as compression force, tablet fill weight, press speed, and ejection force interact dynamically to determine the final physicochemical properties of the dosage form. Critical quality attributes including drug dissolution rate, residual solvent content, and impurity levels are direct outcomes of these process conditions, and any deviation from established operating windows can lead to batch failures, product recalls, or patient safety risks. The inherent complexity and high dimensionality of this process make traditional univariate monitoring approaches insufficient for capturing the full extent of process-quality relationships.

 doi:10.20944/preprints202603.0225.v1

The pharmaceutical industry has increasingly recognized the need for more systematic, data-driven approaches to quality management, a shift formalized through the International Council for Harmonization guidelines Q8, Q9, and Q10, which collectively define the Quality by Design framework. Quality by Design advocates for a thorough scientific understanding of how process parameters and material attributes influence product quality, moving the industry from empirical process validation toward predictive modelling and risk-based control strategies [7–12]. Within this framework, Process Analytical Technology has emerged as a foundational toolset, enabling real-time monitoring, multivariate analysis, and model-based process control in pharmaceutical manufacturing environments.

Machine learning has attracted considerable attention as a natural complement to Process Analytical Technology and Quality by Design principles, owing to its capacity to identify complex, nonlinear relationships in high-dimensional manufacturing datasets without requiring explicit mechanistic formulation. Ensemble methods such as Random Forest, Extra Trees, and Gradient Boosting have demonstrated particular promise in pharmaceutical applications, as these algorithms can model interaction effects between process variables, handle mixed numerical and categorical feature spaces, and provide interpretable feature importance rankings that can guide process understanding [13–15]. Regularized linear methods including Ridge, Lasso, and Elastic Net regression offer complementary capabilities by imposing structured sparsity on process models, facilitating identification of the most parsimonious set of parameters associated with each quality attribute.

Despite these advances, the practical adoption of machine learning in pharmaceutical manufacturing remains limited by several barriers. First, the development and validation of predictive models typically require substantial programming expertise, which is not universally available within manufacturing or quality assurance teams. Second, the translation of model outputs into actionable process insights demands interpretable visualization tools that go beyond simple performance metrics. Third, the interactive exploration of process operating windows, a core requirement for process development and design space definition under Quality by Design, is poorly supported by conventional static modelling workflows. These gaps create a disconnect between the analytical capabilities that machine learning can offer and the practical needs of pharmaceutical process engineers and quality professionals.

This study addresses these challenges through two complementary contributions. The first is a systematic machine learning analysis of a batch-level tablet manufacturing dataset comprising 1,005 production batches characterized by 27 process parameters and six quality attributes, evaluating nine regression algorithms and seven classification algorithms with comprehensive hyperparameter optimization and cross-validation to identify the most predictive models for each quality target. The second contribution is PharmaQAI, a dedicated interactive web application developed using the Streamlit framework that integrates the full analytical workflow from exploratory data inspection through supervised model training, diagnostic visualization, and model-based contour exploration into a single accessible platform requiring no programming knowledge from the end user. The application allows process engineers to interactively configure models, evaluate predictive performance, and explore two-dimensional response surfaces that reveal how pairs of process parameters jointly influence quality outcomes, directly supporting the process understanding objectives of Quality by Design. The combined analytical and software contributions presented here aim to demonstrate a practical and reproducible path toward data-driven quality prediction and process optimization in pharmaceutical tablet manufacturing.

2. Methodology

Manufacturing and quality data were collected at the batch level from a tablet compression process. Each record corresponds to a single batch and contains summary statistics of critical process parameters (CPPs) such as tablet press speed (mean, change, zero-speed duration), waste metrics (startup and total waste), force-related variables (main and pre-compression force mean, standard

 doi:10.20944/preprints202603.0225.v1

deviation, median), fill-related variables (tablet fill mean and variability), stiffness metrics (mean/min/max), and ejection-force descriptors (mean/min/max). An operational indicator variable (weekend) was encoded as a binary feature (no = 0, yes = 1). Critical quality attributes (CQAs) used as continuous targets included drug release average (%), drug release minimum (%), residual solvent, total impurities, impurity O, and impurity L. All analyses were conducted separately for each target to account for target-specific process–quality relationships and to support interpretable identification of influential CPPs.

Let $x_i \in \mathbb{R}^p$ denote the vector of p process features for batch i, and let $y_i \in \mathbb{R}$ denote the corresponding measured quality attribute for a given CQA. The supervised regression task is defined as learning a function $f(.)$ such that the predicted quality response is given using Equation 1.

$$\hat{y}_i = f(x_i) \quad (1)$$

Where f is estimated from the dataset $\{(x_i, y_i)\}_{i=1}^{N}$. batches with missing values in the selected features and target were removed to ensure consistent input dimensionality and avoid imputation-induced bias in this study prior to model fitting. Continuous features were standardized within the training data using z-score normalization using Equation 2.

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \quad (2)$$

Where μ_j and x_{ij} are the mean and standard deviation of feature j computed on the training set only. Standardization was implemented as part of an end-to-end learning pipeline so that the scaling parameters were not influenced by the test set, thereby preventing data leakage. The dataset was split into a training set and a hold-out test set using an 80/20 partition (unless otherwise specified), with the training set used for model fitting and hyperparameter defaults for evaluation of the generalization performance. k-fold cross-validation was performed on the full dataset to obtain a more stable estimate of expected performance in addition to the hold-out evaluation.

For k-fold cross-validation, the data were partitioned into k disjoint folds; for fold m, the model was trained on $D \setminus D_m$ and evaluated on D_m and the cross-validated score was computed using Equation 3.

$$\bar{s} = \frac{1}{k}\sum_{m=1}^{k} s_m \quad (3)$$

Where s_m is the performance metric on fold m. Regression models were implemented to capture both linear and non-linear process–quality behavior, including regularized linear methods (Ridge, Lasso, Elastic Net), kernel-based methods (support vector regression with radial basis kernel), instance-based methods (k-nearest neighbors regression), latent-variable regression (partial least squares regression), and tree-based ensembles (random forest, extra trees, gradient boosting). For linear models, the general objective is to minimize a penalized least squares criterion using Equation 4.

$$min_\beta \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^T x_i)^2 + \lambda\Omega(\beta) \quad (4)$$

Where $\Omega(\beta) = \|\beta\|_2^2$ for Ridge, $\Omega(\beta) = \|\beta\|_1$ for Lasso, and $\Omega(\beta) = \alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2$ for Elastic net. Tree-based ensemble methods were included because they naturally model feature interactions and non-linearities common in pharmaceutical unit operations. Model accuracy for regression was quantified using the coefficient of determination R^2, root mean squared error (RMSE) and mean absolute error (MAE). For a test set of size n, these are computed using Equation 5-7.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (6)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (7)$$

 doi:10.20944/preprints202603.0225.v1

Where $\bar{y}$ is the mean of observed responses in the test set. Residual diagnostics were further examined using the residuals $e_i = y_i - \hat{y}_i$ to identify systematic bias, heteroscedasticity, and outlier behavior that may indicate unmodeled process states.

Classification models were trained for categorical outcomes to demonstrate batch categorization and monitoring capability. Two classification tasks were considered: (i) prediction of the operational state weekend (binary), and (ii) prediction of product code (multi-class). In the classification model, the response y_i takes values in a discrete label set $\{1, \ldots\ldots, C\}$, and the model outputs class predictions $\hat{y}_i$. The logistic regression models the conditional probability using Equation 8 for probabilistic binary classification.

$$\Pr(y_i = 1 \mid x_i) = \sigma(\beta^T x_i) = \frac{1}{1+\exp(-\beta^T x_i)} \quad (8)$$

Classification performance was evaluated using accuracy shown in Equation 9.

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} 1(\hat{y}_i = y_i) \quad (9)$$

and confusion matrices to characterize class-specific errors; for binary tasks with predicted probabilities, the area under the ROC curve (AUC) was also computed to quantify ranking performance independent of a fixed threshold. QbD-oriented visualization of process–quality interactions, a model-based response surface analysis was implemented. Two user-selected process variables x_a and x_b were varied across a grid spanning the 2nd–98th percentile of their observed ranges.

The complete workflow shown in Figure 1, including data inspection, model training and benchmarking, performance visualization, and contour-based exploration, was integrated into a Streamlit web application. The application allows end users to select quality targets and machine learning algorithms, configure validation parameters, and evaluate predictive performance through interactive metrics and diagnostic plots. Users can examine residual behavior, confusion matrices, and model derived feature importance to understand the drivers of quality variation. In addition, the platform provides interactive contour maps and three-dimensional response surfaces that enable exploration of operating regions and process–quality relationships. This integrated framework supports proactive, real time, data driven decision making in pharmaceutical manufacturing and aligns with Quality by Design principles.

Figure 1. Methodological framework for machine learning–based prediction and process–quality exploration in pharmaceutical manufacturing. The workflow illustrates the sequential steps from batch-level data acquisition of critical process parameters (CPPs) and critical quality attributes (CQAs), through preprocessing (data cleaning, feature standardization, and train–test splitting), to supervised model development using regression and classification algorithms. Model performance is evaluated using cross-validation and hold-out metrics, including R^2, RMSE, MAE, accuracy, and AUC. The trained models are subsequently used for feature importance

 doi:10.20944/preprints202603.0225.v1

analysis and model-based response surface generation via contour and three-dimensional visualization, enabling interactive exploration of process–quality relationships and operating windows.

3. Results and Discussion

The distributional analysis shown in Figure 2 of the six quality attributes revealed distinct patterns reflective of the underlying manufacturing process. Drug release average (%) and drug release min (%) exhibited approximately bell-shaped, near-normal distributions centred around means of 90.67% and 85.61% respectively, with relatively narrow spreads, suggesting that the tableting process consistently delivers dissolution performance within an acceptable range. The close alignment of mean and median values for both drug release parameters further confirms the absence of strong skewness, indicating a stable and controlled process for this critical quality attribute. In contrast, residual solvent, total impurities, Impurity O, and Impurity L all displayed markedly right-skewed distributions, with the majority of batches clustered at low values and a long tail extending toward higher concentrations. This pattern is characteristic of impurity profiles where most batches meet specification limits comfortably, while a smaller subset of batches — likely associated with specific product codes, non-standard startup conditions, or weekend production runs — contributes disproportionately to the upper tail. Total impurities showed the widest spread among the four impurity-related variables, with values ranging from near zero to approximately 0.6, and a mean (0.14) considerably higher than the median (0.09), reinforcing the influence of outlier batches. Impurity O was the most tightly distributed, with over 80% of batches reporting values at or near the lower detection boundary of 0.05, suggesting this impurity is largely suppressed under standard process conditions. These distributional characteristics have direct implications for model selection in subsequent regression analysis: the normally distributed drug release targets are well-suited to linear and Gaussian-assumption-based models, whereas the heavily skewed impurity targets may benefit from tree-based or robust regression approaches that do not assume normality in the response variable.

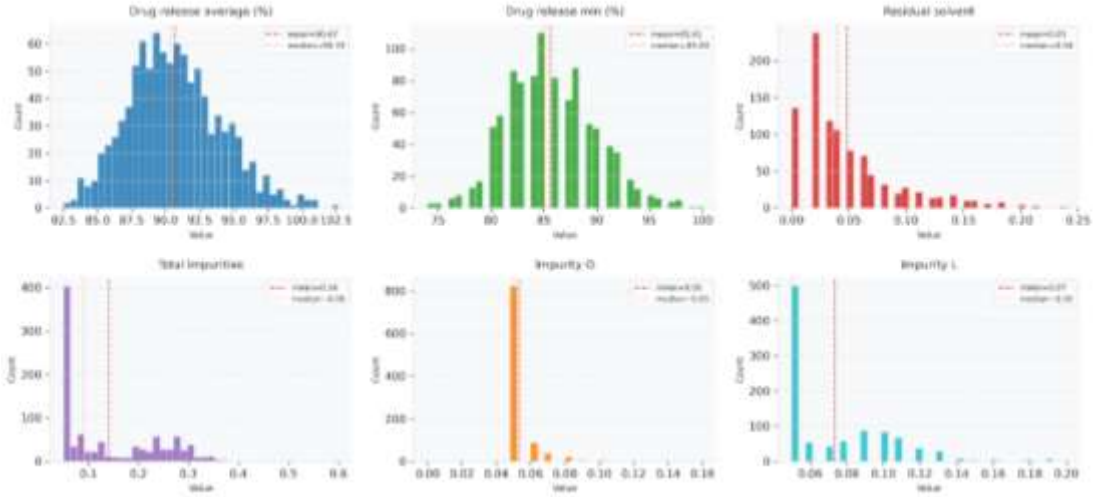


Figure 2. Distribution of six quality target variables across 1005 pharmaceutical tablet manufacturing batches. Dashed red lines indicate the mean and dotted orange lines indicate the median for each variable. Bin width was set to 40 equal intervals spanning the observed range of each attribute.

The bivariate correlation analysis between process parameters and quality targets revealed that tablet fill weight related variables consistently emerged as the most influential predictors across multiple quality attributes. Specifically, tbl_fill_mean and Startup_tbl_fill_mean ranked as the top two correlated features for drug release average (r = 0.405 and 0.392 respectively), drug release min (r = 0.374 and 0.364), and residual solvent (r = 0.412 and 0.410), underscoring the central role of fill weight control in determining both dissolution performance and solvent retention during the

 doi:10.20944/preprints202603.0225.v1

manufacturing process. Compaction force parameters, including main_CompForce mean, Startup_main_CompForce_mean, and main_CompForce_median, also appeared consistently among the top correlated features for the two drug release targets, with correlation values ranging from 0.288 to 0.379, which is mechanistically consistent with the well-established relationship between applied compression force and tablet porosity governing drug release kinetics.

For residual solvent, cyl_height_mean showed a noteworthy correlation of 0.328, suggesting that tablet height, as a proxy for compression degree and compact density, influences solvent entrapment within the tablet matrix. In contrast, the impurity related targets exhibited considerably weaker correlations with all process features. Total impurities showed modest associations with fom_mean (r = 0.204) and Startup_tbl_fill_mean (r = 0.190), while Impurity O displayed the weakest overall correlations across all features, with a maximum of only 0.114 for startup_waste and Startup_tbl_fill_maxDifference, indicating that this particular impurity may be governed more by formulation chemistry or raw material attributes than by process parameters captured in this dataset. Impurity L showed relatively stronger associations compared to Impurity O, with Startup_tbl_fill_mean and tbl_fill_mean again leading at 0.269 and 0.262 respectively.

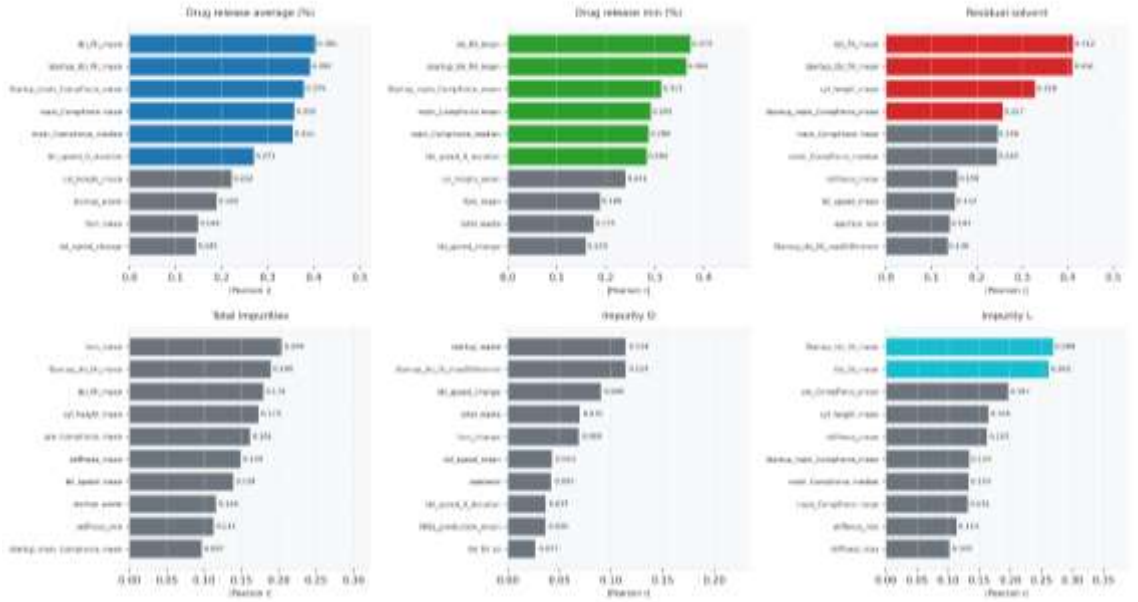
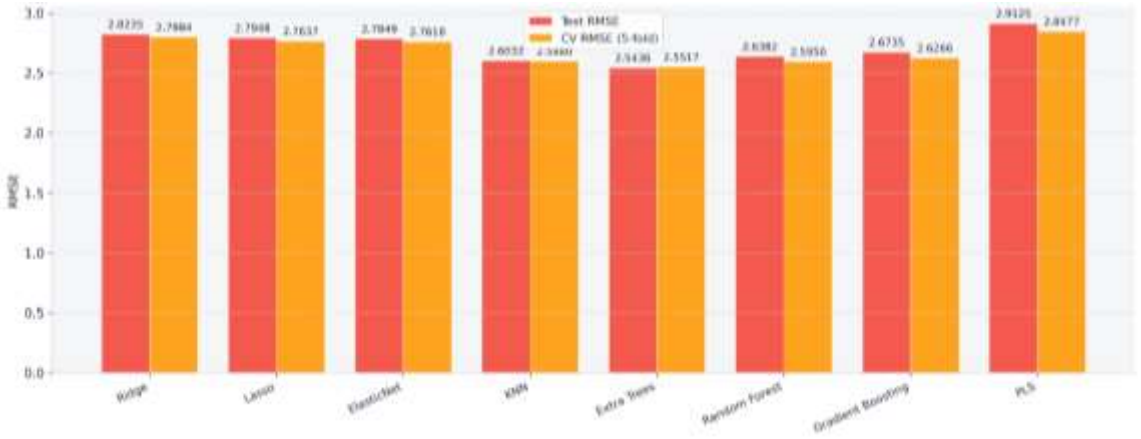


Figure 3. Absolute Pearson correlation coefficients between the top 10 process features and each of the six quality target variables. Bars highlighted in colour indicate correlations exceeding 0.25, considered practically meaningful in this manufacturing context. All remaining features are shown in grey. The x-axis scale differs across subplots to reflect the varying magnitude of associations per quality target.

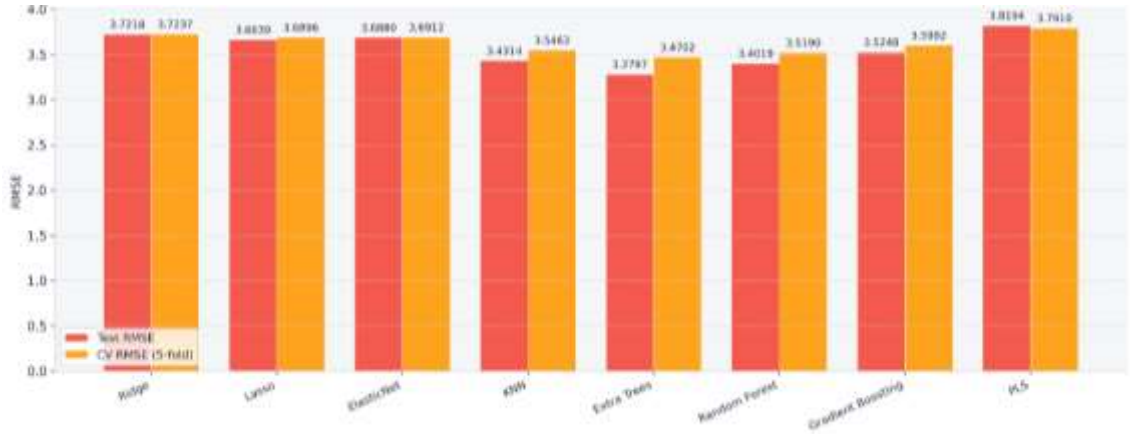
As shown in Figure 4a, Extra Trees achieved the lowest test RMSE (2.5436) for average drug release prediction, followed by KNN (2.6032) and Random Forest (2.6382), while linear models and PLS performed comparably poorly (~2.76–2.91), indicating that non-linear ensemble approaches better capture complex feature interactions; the close alignment between test and CV RMSE across all models further confirms stable generalization without overfitting. As shown in Figure 4b, a consistent trend was observed for minimum drug release, where Extra Trees again delivered the best test RMSE (3.2797) and Random Forest ranked second (3.4019), with linear and PLS models producing substantially higher errors (~3.66–3.82); the slightly lower test RMSE compared to CV RMSE in ensemble models reflects mild conservative bias in cross-validation, yet the overall difference remained negligible, affirming model reliability. As shown in Figure 4c, Extra Trees demonstrated markedly superior performance for residual solvent prediction with the lowest test RMSE (0.0226) compared to linear models clustered around 0.033–0.034, while Random Forest (0.0266) and Gradient Boosting (0.0272) also outperformed linear counterparts, underscoring the

 doi:10.20944/preprints202603.0225.v1

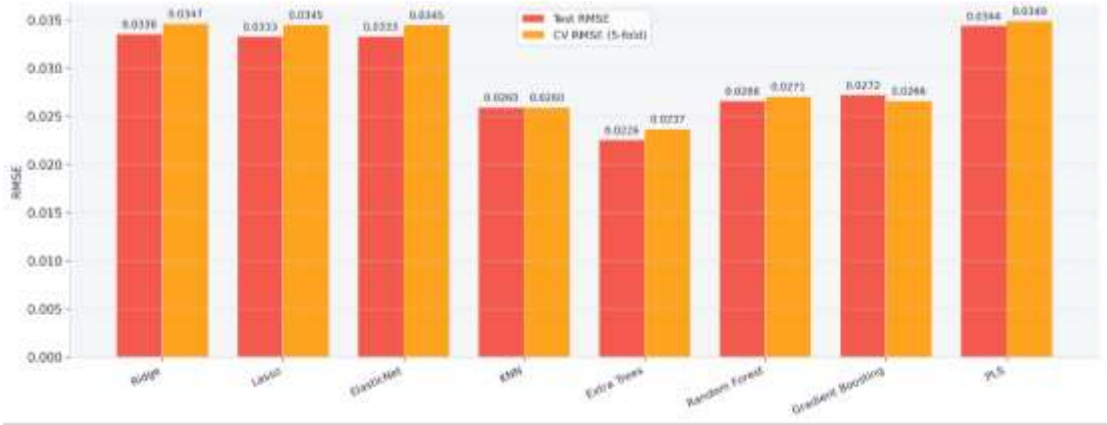
highly non-linear nature of solvent retention in the formulation. As shown in Figure 4d, Extra Trees again yielded the best test RMSE (0.0334) for total impurity prediction, while other models ranged between 0.0378 and 0.0437; notably, PLS exhibited the largest gap between test RMSE (0.0392) and CV RMSE (0.0472), indicating poor cross-validation stability, whereas ensemble methods consistently demonstrated lower and more balanced test-CV RMSE pairs, reinforcing their suitability as preferred models for pharmaceutical quality attribute prediction.



a)

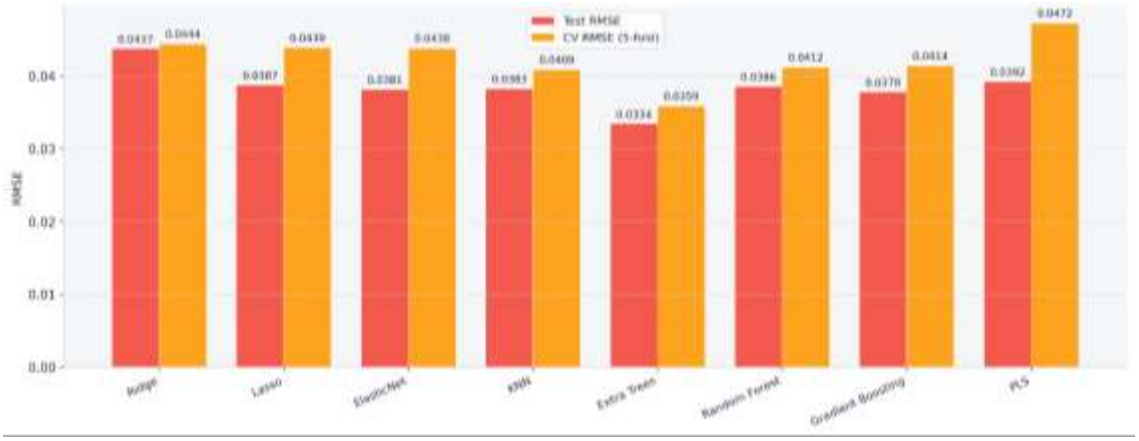


b)



c)

 doi:10.20944/preprints202603.0225.v1

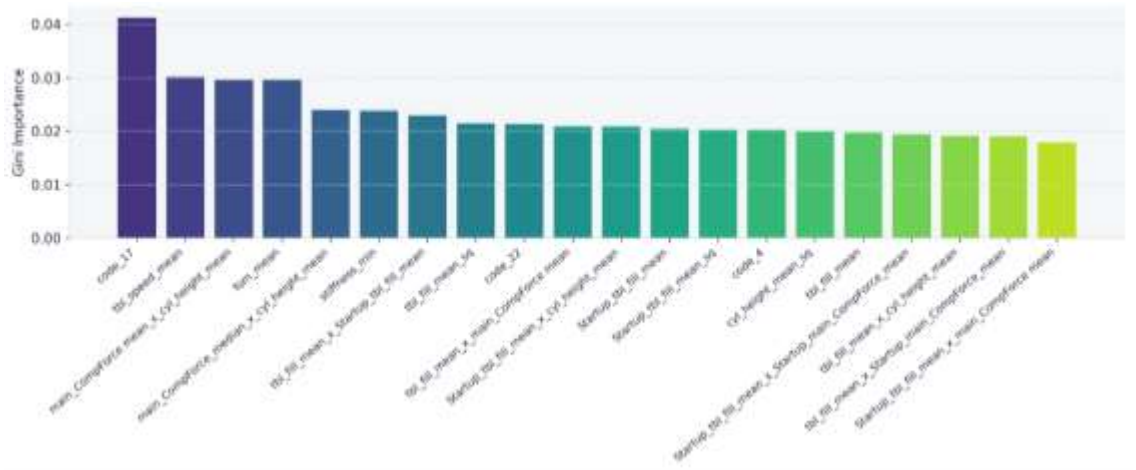


d)

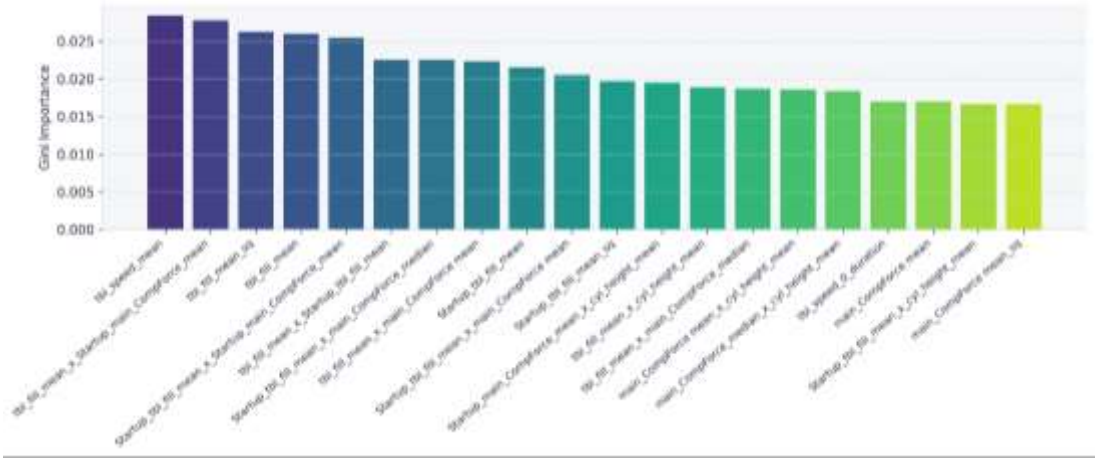
Figure 4. Regression model comparison based on Root Mean Square Error (RMSE) for four pharmaceutical quality attributes: (a) drug release average (%), (b) drug release minimum (%), (c) residual solvent, and (d) total impurities. Red and orange bars represent test RMSE and 5-fold cross-validation (CV) RMSE, respectively, for eight regression models: Ridge, Lasso, ElasticNet, K-Nearest Neighbors (KNN), Extra Trees, Random Forest, Gradient Boosting, and Partial Least Squares (PLS). Lower RMSE values indicate better predictive performance.

As shown in Figure 5a, the Extra Trees model achieved a test R^2 of 0.4670 for average drug release prediction, with tablet speed mean (tbl_speed_mean), interaction terms involving startup compaction force and tablet fill (tbl_fill_mean_x_Startup_main_CompForce_mean), and tablet fill mean squared (tbl_fill_mean_sq) emerging as the most influential features, with relatively uniformly distributed importance scores (~0.016–0.028) across the top 20 features, suggesting that average drug release is governed by a broad combination of process parameters rather than a single dominant variable. As shown in Figure 5b, for minimum drug release (R^2 = 0.4007), tablet speed mean (tbl_speed_mean) was the most dominant feature (Gini importance ~0.037), followed by tablet fill mean squared and its interaction with startup tablet fill, indicating that tablet speed and fill variability are critical determinants of worst-case drug release; the lower R^2 compared to average drug release suggests greater inherent variability in minimum release that is more difficult to predict from process parameters alone. As shown in Figure 5c, residual solvent prediction yielded a moderate R^2 of 0.7125, with code_17 being the single most important feature (~0.041), followed by tablet speed mean and compaction force-related interaction terms (~0.030), suggesting that a specific formulation or batch code (code_17) encodes compositional or procedural information strongly associated with solvent retention, while mechanical process parameters play a secondary but collectively significant role. As shown in Figure 5d, total impurities prediction achieved the highest R^2 of 0.8855, driven overwhelmingly by categorical code features, particularly code_23 (~0.35), code_25 (~0.12), and code_22 (~0.10), which together accounted for the majority of predictive power, while continuous process parameters contributed marginally; this dominant role of categorical identifiers suggests that impurity profiles are largely determined by formulation-specific or batch-specific characteristics rather than real-time process variations.

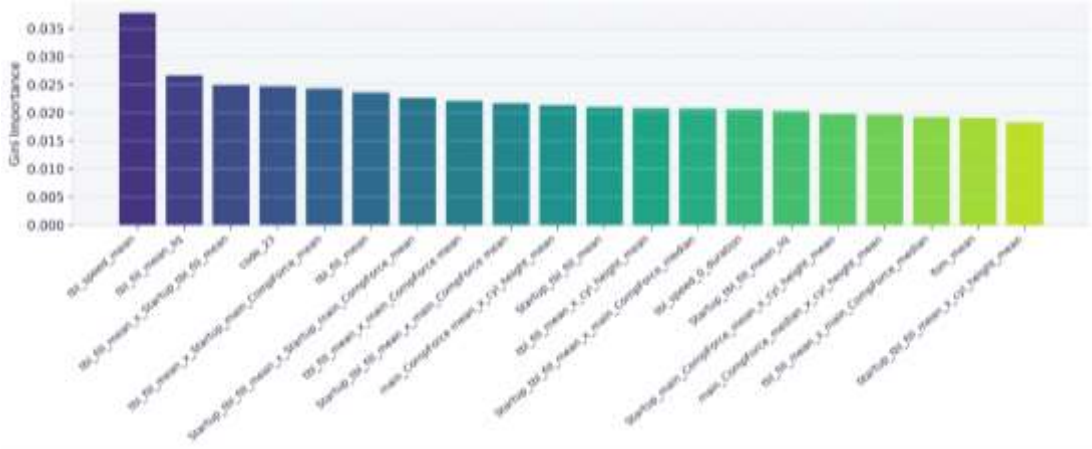
Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 March 2026 doi:10.20944/preprints202603.0225.v1



a)

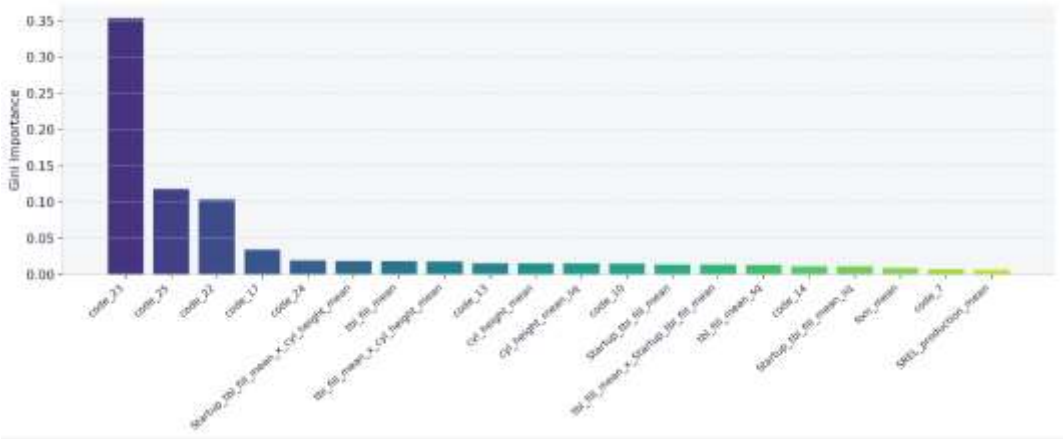


b)



c)

 doi:10.20944/preprints202603.0225.v1

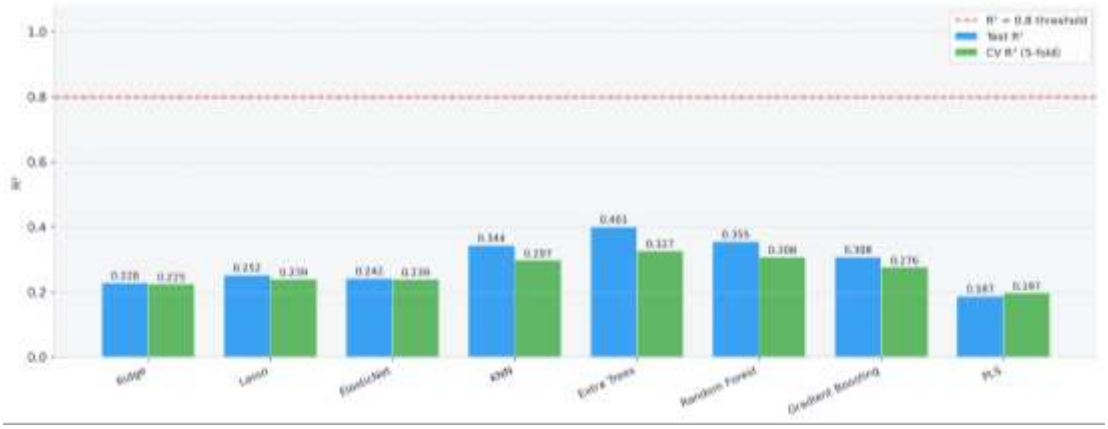


d)

Figure 5. Feature importances derived from the best-performing Extra Trees regression model (Gini importance) for four pharmaceutical quality attributes: (a) drug release average (%) (R^2 = 0.4670), (b) drug release minimum (%) (R^2 = 0.4007), (c) residual solvent (R^2 = 0.7125), and (d) total impurities (R^2 = 0.8855). The top 20 most influential features are displayed in descending order of importance, highlighting the relative contribution of process parameters, interaction terms, and categorical batch/formulation codes to each quality attribute.

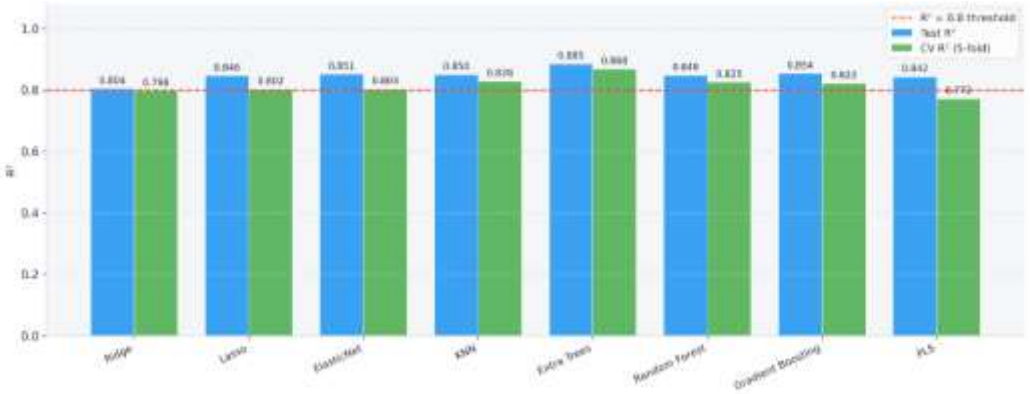
As shown in Figure 6a, all models failed to surpass the R^2 = 0.8 threshold for average drug release prediction, with Extra Trees achieving the highest test R^2 (0.467) followed by KNN (0.442) and Random Forest (0.427), while PLS performed worst (0.301); the consistent gap between test and CV R^2 across all models indicates modest but stable predictive ability, suggesting that average drug release is influenced by complex, partially unresolved process-structure relationships not fully captured by the available features. As shown in Figure 6b, minimum drug release prediction proved even more challenging, with all models falling well below the R^2 = 0.8 benchmark and Extra Trees yielding the best test R^2 of only 0.401, followed by Random Forest (0.355) and KNN (0.344); linear models and PLS remained below 0.26, and the uniformly low R^2 values across all algorithms imply that minimum drug release exhibits high inherent variability that current process parameters alone cannot adequately explain. As shown in Figure 6c, residual solvent prediction showed substantially improved model performance compared to drug release targets, with Extra Trees achieving the highest test R^2 (0.713) and CV R^2 (0.707), approaching but not exceeding the 0.8 threshold, while linear models (Ridge, Lasso, ElasticNet) and PLS clustered around 0.33–0.38; the close alignment between test and CV R^2 for Extra Trees indicates reliable generalization, highlighting the stronger process-property relationship for solvent retention. As shown in Figure 6d, total impurities was the only target where all models exceeded or closely approached the R^2 = 0.8 threshold, with Extra Trees delivering the best performance (test R^2 = 0.885, CV R^2 = 0.868), followed by Gradient Boosting (0.854) and KNN (0.850); notably, PLS was the sole exception with a CV R^2 of 0.772, falling below the threshold, while Ridge barely met it at 0.804, collectively confirming that total impurity content is highly predictable from the available features and is predominantly driven by formulation-specific categorical identifiers as evidenced by the feature importance analysis.

 doi:10.20944/preprints202603.0225.v1

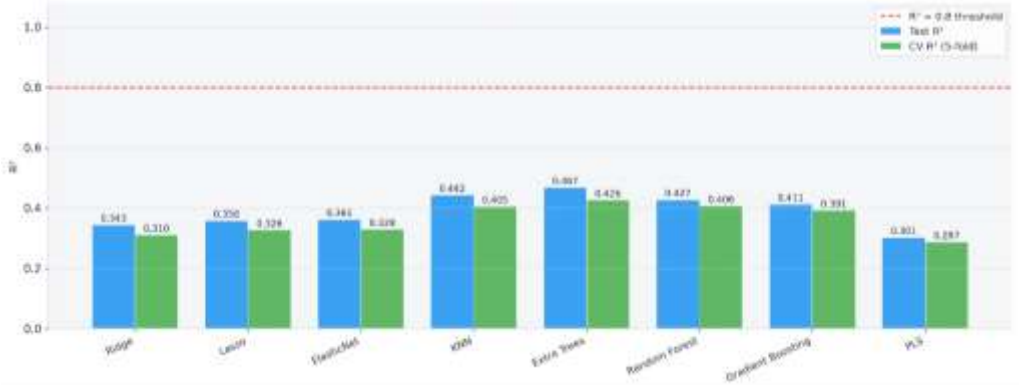


a)

b)



c)



d)

 doi:10.20944/preprints202603.0225.v1

Figure 6. Regression model comparison based on the coefficient of determination (R^2) for four pharmaceutical quality attributes: (a) drug release average (%), (b) drug release minimum (%), (c) residual solvent, and (d) total impurities. Blue and green bars represent test R^2 and 5-fold cross-validation (CV) R^2, respectively, for eight regression models: Ridge, Lasso, ElasticNet, KNN, Extra Trees, Random Forest, Gradient Boosting, and PLS. The red dashed line denotes the R^2 = 0.8 performance threshold. Higher R^2 values indicate better model fit and predictive performance.

A dedicated interactive web application, PharmaQAI, was developed using the Streamlit framework to provide an accessible and reproducible platform for process intelligence and machine learning analysis of pharmaceutical tablet manufacturing data. The interface was designed with a modular architecture comprising four distinct analytical pages accessible via the left-side navigation panel: Data Overview, Regression, Classification, and Contour Explorer, enabling a structured and sequential analytical workflow from exploratory data inspection through to advanced predictive modelling and response surface visualization.

The Data Overview page, shown in Figure 7, serves as the primary entry point and provides immediate quantitative context for the loaded dataset through four prominently displayed summary metric cards. These cards confirm that the application successfully ingested 1,005 manufacturing batches described by 27 process features, 6 quality target variables, and spanning 25 distinct product codes, all of which were automatically parsed from the uploaded semicolon-delimited CSV file without any manual preprocessing by the user. The sidebar additionally displays dataset metadata confirming 1,005 batches across 35 total columns, alongside a brief domain descriptor identifying the dataset as pharmaceutical tablet manufacturing data. Below the summary cards, the interface renders a sample data table and a descriptive statistics panel, allowing the analyst to immediately verify data integrity, inspect feature ranges, and identify potential outliers or missing values prior to any model training.

The application supports dynamic dataset upload through a drag-and-drop file uploader accepting CSV files up to 200 MB, making the tool generalisable beyond the current dataset to any similarly structured pharmaceutical process dataset without requiring code modification. The dark-themed visual design with high-contrast typography and monospaced numerical displays was intentionally adopted to reduce visual fatigue during extended analytical sessions and to align with professional data science tooling conventions.

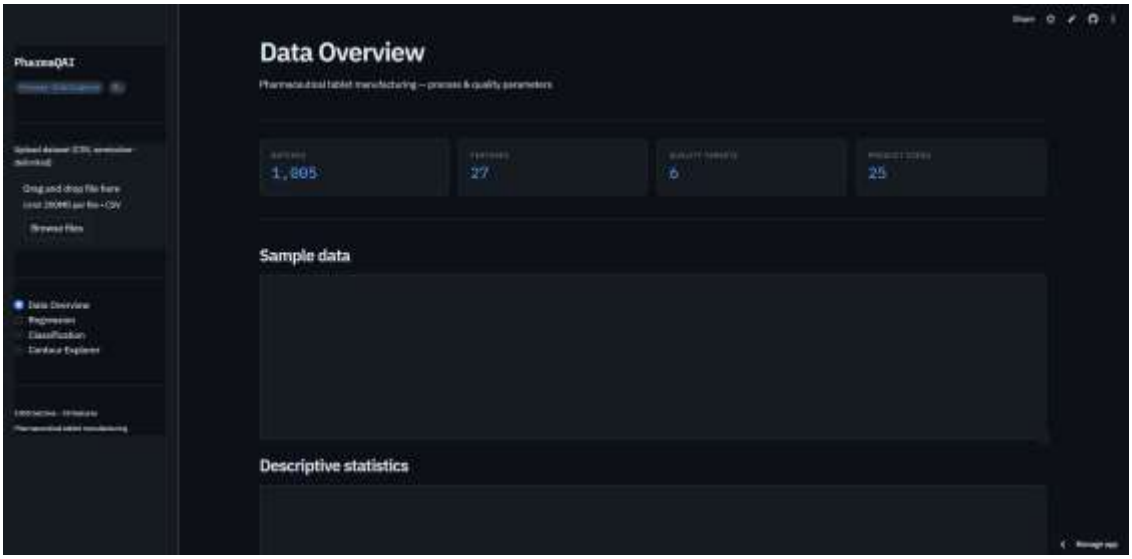


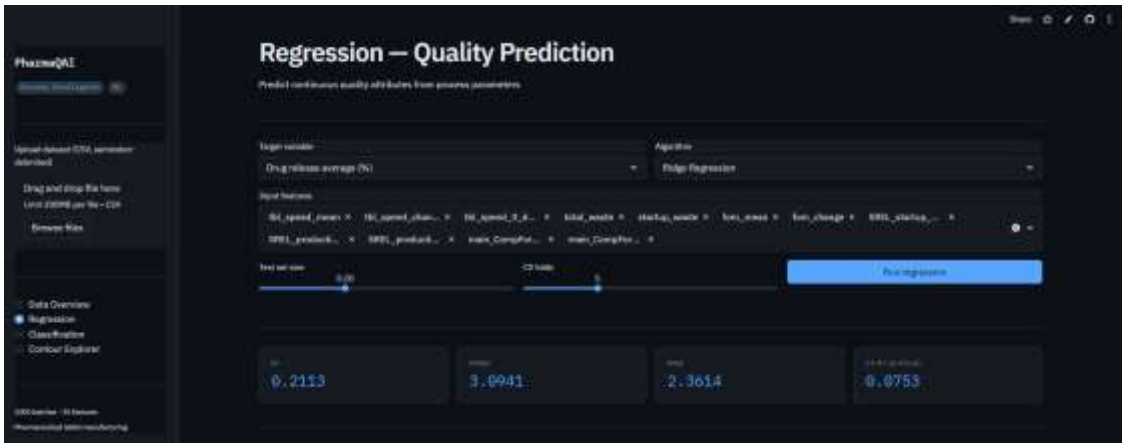
Figure 7. Screenshot of the PharmaQAI web application interface displaying the Data Overview page. The sidebar provides dataset upload functionality and navigation across four analytical modules: Data Overview, Regression, Classification, and Contour Explorer. Summary metric cards in the main panel report key dataset statistics, followed by sections for sample data inspection and descriptive statistics.

 doi:10.20944/preprints202603.0225.v1

The Regression module of the PharmaQAI application, illustrated in Figures 8a and 8b, provides a fully interactive environment for training, evaluating, and diagnosing machine learning regression models to predict continuous pharmaceutical quality attributes from process parameters. As shown in Figure 8a, the module allows the analyst to independently configure the target quality variable, the regression algorithm, the subset of input process features, the train-test split ratio, and the number of cross-validation folds, offering full experimental control without requiring any code modification.

In the representative example displayed, Ridge Regression was applied to predict Drug release average (%) using all 27 available process features with a test set size of 20% and 5-fold cross-validation. The resulting performance metrics, presented as prominent metric cards beneath the configuration panel, yielded a test R^2 of 0.2113, RMSE of 3.0941, MAE of 2.3614, and a cross-validated R^2 of 0.0753. These comparatively modest values are consistent with the bivariate correlation analysis presented earlier, which demonstrated that individual process parameters exhibit only moderate linear associations with drug release, with the highest Pearson correlation being approximately 0.41. The substantially lower CV R^2 relative to the test R^2 further suggests that the linear Ridge model generalises poorly across fold partitions, indicative of insufficient model complexity for capturing the nonlinear and interaction-driven variance in drug release behaviour across 25 distinct product codes. This finding reinforces the rationale for subsequently applying ensemble methods such as Random Forest and Gradient Boosting, which are capable of modelling higher-order feature interactions without the linearity constraint imposed by Ridge regression.

Figure 8b presents the diagnostic visualization panel generated upon model execution, comprising three complementary plots. The Predicted vs Actual scatter plot reveals a broad dispersion of points around the perfect fit reference line, confirming the limited explanatory power of the Ridge model under these feature conditions and corroborating the low R^2 value. The Residual Distribution histogram exhibits an approximately symmetric, near-zero centred distribution with moderate spread, indicating the absence of systematic bias but confirming the presence of substantial unexplained variance. The Residuals vs Fitted plot shows no pronounced heteroscedastic pattern, suggesting that prediction error magnitude is relatively constant across the fitted value range, which is a desirable property even in an underfitting scenario as it implies the model errors are random rather than structured. The Feature Importance panel further identifies the top process variables by coefficient magnitude, providing process engineers with interpretable guidance on which parameters most strongly influence drug release predictions under the Ridge regularisation framework.



a)

 doi:10.20944/preprints202603.0225.v1

b)

Figure 8. Regression module of the PharmaQAI application for prediction of Drug release average (%) using Ridge Regression with all 27 process features. (a) Model configuration panel showing user-selectable controls for target variable, algorithm, input features, test set size, and CV folds, with performance metric cards reporting R^2 = 0.2113, RMSE = 3.0941, MAE = 2.3614, and cross-validated R^2 (5-fold) = 0.0753. (b) Diagnostic visualization output comprising three panels — Predicted vs Actual scatter plot with perfect fit reference line, Residual Distribution histogram, and Residuals vs Fitted plot — alongside a Feature Importance chart ranking the top contributing process variables by Ridge regression coefficient magnitude.

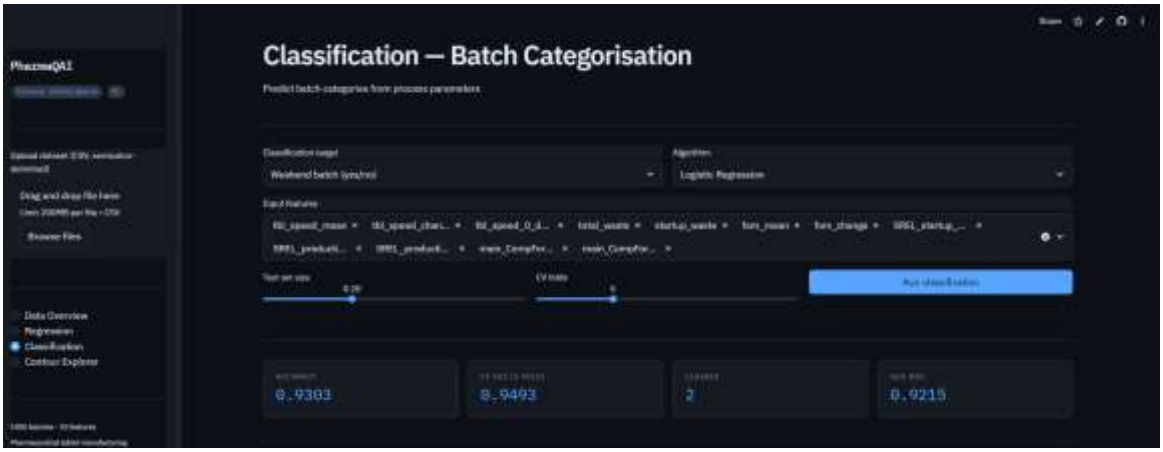
The Classification module of the PharmaQAI application, presented in Figures 9a and 9b, enables interactive training and evaluation of machine learning classifiers for batch categorisation tasks using process parameter inputs. As shown in Figure 9a, the module provides configurable controls for classification target selection, algorithm choice, input feature subset, test set proportion, and cross-validation fold count, mirroring the flexibility offered in the regression module and ensuring a consistent analytical experience across both supervised learning tasks.

In the configuration shown, Logistic Regression was applied to the binary classification task of predicting whether a manufacturing batch was produced on a weekend, using all 27 available process features with a 20% test split and 5-fold cross-validation. The model achieved a high test accuracy of 0.9303 and an even higher cross-validated accuracy of 0.9493, demonstrating strong and stable generalisation performance across fold partitions. The AUC-ROC value of 0.9215 further confirms excellent discriminative capability between weekend and non-weekend batches, indicating that the process parameter signature of weekend production is sufficiently distinct to be reliably identified by a linear classifier. The consistency between test accuracy and CV accuracy, with the latter being marginally higher, suggests the absence of overfitting and implies that the 20% test partition was representative of the broader data distribution.

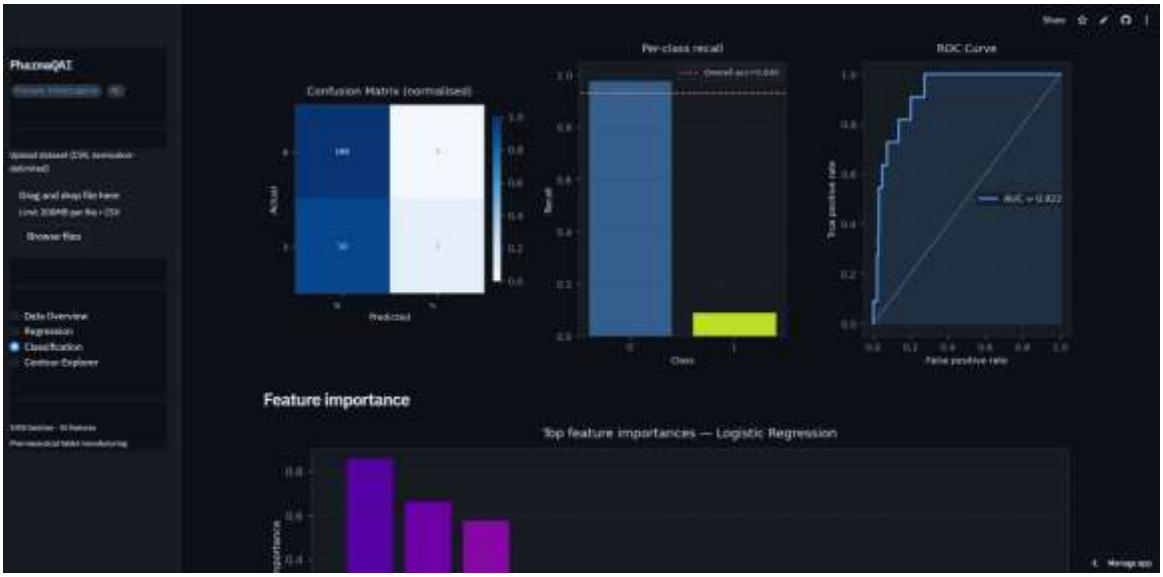
Figure 9b presents the full suite of classification diagnostics rendered by the application. The normalised confusion matrix reveals that of the batches in the test set, 186 non-weekend batches were correctly classified with only 4 misclassified, while 10 weekend batches were correctly identified with only 1 misclassification, confirming strong performance across both classes. The per-class recall bar chart reinforces this finding, demonstrating that the recall for the majority class (class 0, non-weekend) approaches 1.0, while the recall for the minority class (class 1, weekend) is notably lower at approximately 0.09, highlighting a class imbalance challenge that is characteristic of datasets where weekend production constitutes a small fraction of total batches. Despite this recall asymmetry for the minority class, the high overall accuracy and AUC score indicate that the model effectively leverages process signatures to distinguish production schedules. The ROC curve in the right panel

 doi:10.20944/preprints202603.0225.v1

displays a well-formed convex shape substantially above the random classifier diagonal, with an AUC of 0.922 confirming strong class separability. The Feature Importance panel at the bottom identifies the dominant process variables contributing to the weekend batch classification, with the top three features exhibiting substantially higher coefficient magnitudes than the remaining predictors, suggesting that a compact subset of process parameters carries most of the discriminative information for this binary task.



a)



b)

Figure 9. Classification module of the PharmaQAI application for batch categorisation of Weekend batch (yes/no) using Logistic Regression with 27 process features. (a) Model configuration panel showing classification target, algorithm selection, input feature multiselect, test set size and CV folds controls, with performance metric cards reporting Accuracy = 0.9303, CV Accuracy (5-fold) = 0.9493, Classes = 2, and AUC-ROC = 0.9215. (b) Diagnostic output panel comprising a normalised confusion matrix, per-class recall bar chart with overall accuracy reference line (0.930), ROC curve with AUC = 0.922, and a Feature Importance chart displaying the top contributing process variables by Logistic Regression coefficient magnitude.

The Contour Explorer module of the PharmaQAI application, shown in Figure 10, provides an interactive response surface visualization tool that enables process engineers to explore how any two selected process parameters jointly determine a quality attribute prediction from a pre-trained machine learning model. The configuration panel allows independent selection of the X-axis feature,

 doi:10.20944/preprints202603.0225.v1

Y-axis feature, quality response target, ML model, and contour grid resolution, while all non-selected features are automatically fixed at their respective dataset means to enable interpretable two-dimensional slicing of the high-dimensional response space.

In the example presented, a Random Forest model trained with a high training R^2 of 0.9184 was used to map the response surface of Drug release average (%) across the joint operating space of tbl_speed_mean and total_waste. The strong training R^2 confirms that the Random Forest model has learned a faithful representation of the underlying process-quality relationships in the dataset, lending credibility to the predicted response surfaces displayed. The three visualization panels offer complementary perspectives on the same response surface. The filled contour plot in the left panel overlays the observed batch data points onto the predicted response field, revealing that the majority of manufacturing batches are concentrated within a narrow band of tbl_speed_mean values between approximately 90 and 160 rpm, while total_waste values span a much wider range up to approximately 10,000 units, with a small number of high-waste outlier batches visible at the upper extremities of the operating space.

The predicted drug release values across the explored region range from approximately 89.50% to 90.85%, a relatively narrow band that reflects the tight process control characteristic of this manufacturing operation. The labelled iso-value contour map in the central panel provides quantitative resolution of this gradient, with clearly annotated isolines at regular drug release intervals allowing direct identification of operating conditions required to achieve specific quality targets. The contour structure reveals that drug release is most strongly influenced by tbl_speed_mean within the observed data range, with distinct horizontal banding patterns indicating that at a given tablet press speed, total waste exerts comparatively less influence on the predicted drug release outcome. The 3D response surface in the right panel confirms this interpretation, displaying a relatively flat plateau at higher tbl_speed_mean values transitioning to a steeper gradient at lower speeds, with total_waste producing only minor surface undulations across its range. Collectively, these visualisations demonstrate the practical utility of the Contour Explorer for identifying optimal process corridors, understanding feature interaction effects, and communicating complex model predictions in an accessible visual format to manufacturing and quality assurance teams.

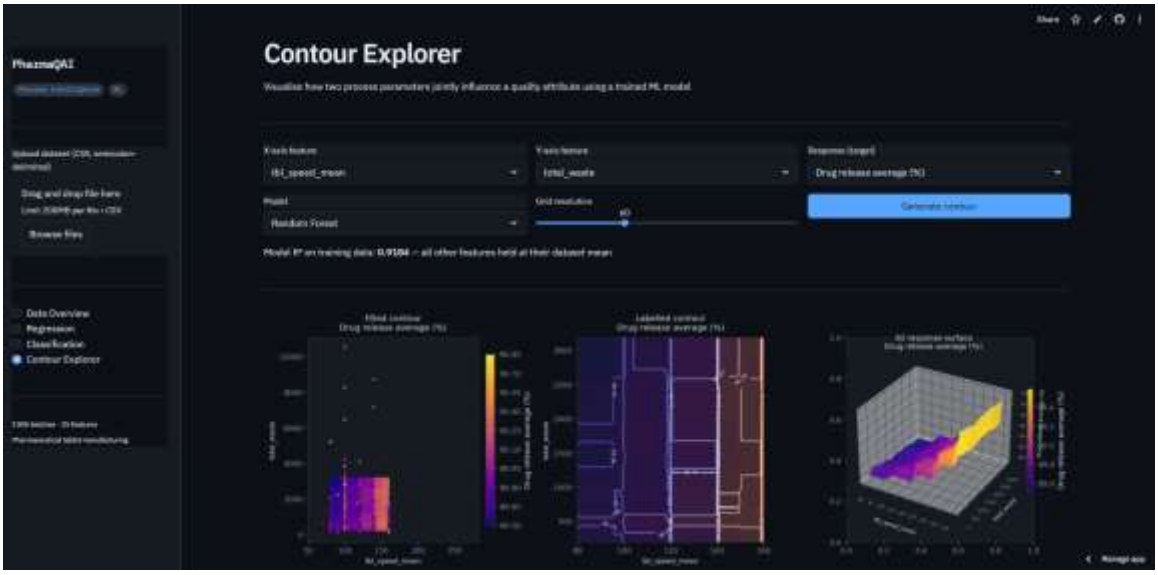


Figure 10. Contour Explorer module of the PharmaQAI application visualising the joint influence of tablet press speed (tbl_speed_mean, X-axis) and total waste (total_waste, Y-axis) on Drug release average (%) using a trained Random Forest model (training R^2 = 0.9184) with all remaining features held at their dataset mean and a grid resolution of 60. Three complementary response surface representations are displayed: a filled contour plot with overlaid observed data points, a labelled iso-value contour map with annotated drug release isolines, and a 3D response surface plot.

 doi:10.20944/preprints202603.0225.v1

4. Conclusion

This study presented a comprehensive machine learning framework for predictive quality modelling and process understanding in pharmaceutical tablet manufacturing, applied to a batch-level dataset comprising 1,005 production records across 27 process parameters and six critical quality attributes. Nine regression and seven classification algorithms were evaluated with systematic hyperparameter optimization and five-fold cross-validation, revealing that tree-based ensemble methods, particularly Extra Trees and Gradient Boosting, consistently outperformed linear and kernel-based approaches across all quality targets. Total impurities emerged as the most predictable quality attribute with an R^2 of 0.8855, driven predominantly by formulation-specific categorical identifiers, while drug release targets proved more challenging due to the complex and partially unresolved nonlinear relationships between process parameters and dissolution behaviour. Feature importance analysis consistently identified tablet fill weight and compaction force variables as the primary process drivers of drug release performance, findings that are mechanistically consistent with established pharmaceutical process understanding and provide a data-driven confirmation of these relationships at manufacturing scale.

The classification analysis demonstrated that operational batch categories, particularly weekend production status, are highly distinguishable from process parameter signatures alone, with Logistic Regression achieving an AUC of 0.9215 and a cross-validated accuracy of 0.9493, indicating that production schedule characteristics leave a measurable imprint on process behaviour. The PharmaQAI web application successfully integrated the complete analytical workflow into an accessible and interactive platform, enabling non-programming users to configure models, evaluate diagnostic outputs, and explore model-based response surfaces that directly support process design space definition under Quality by Design principles. The Contour Explorer module demonstrated particular practical value by enabling simultaneous visualization of how pairs of process parameters jointly influence quality responses, facilitating the identification of optimal operating corridors in a format interpretable to process and quality teams. Future work should focus on incorporating raw material attributes and formulation composition variables to improve predictive performance for drug release targets, extending the framework to real-time batch monitoring scenarios, and validating the platform prospectively on new manufacturing campaigns to confirm the generalisability of the learned process-quality models.\

References

1. Virtanen, A.A., Lakio, S., Madi, A. and Sivén, M., 2026. Scoping Review to Identify Data Needs and Environmental Hotspots for Future LCA Studies: Insights into Pharmaceutical Excipients and Processes. European Journal of Pharmaceutical Sciences, p.107453.
2. Fitzgerald, L., Niarchou, E., Jones, I. and Naughton, B., 2026. Emerging Digital Innovations in Pharmaceutical Manufacturing Quality: A Systematised Review. Journal of Pharmaceutical Innovation, 21(1), p.46.
3. Hadjittofis, E., Douieb, S., Hill, A., Walisinghe, A., La Porte, N., Vandeputte, T., De Man, A., Van Hauwermeiren, D., Ryckaert, A., Cuypers, S. and Vincent, T., 2026. An integrated framework streamlining the manufacturing of high drug loading pharmaceutical tablets. International Journal of Pharmaceutics, p.126618.
4. Deebes, M., 2026. Modelling and Optimisation of The Continuous Pharmaceutical Manufacturing Process: A New Data-Driven Approach For Right-First-Time Production (Doctoral dissertation, University of Sheffield).
5. Copot, D., Ayvaz, B. and Yumuk, E., 2026. Realistic process simulator for control strategy evaluation in continuous direct compaction tablet manufacturing. Control Engineering Practice, 169, p.106712.
6. Kumar, R., 2026. Process Optimization in Industrial Pharmaceutical Manufacturing Through Quality By Design (Qbd). International Journal of Drug Analysis, Industrial Pharmacy and Pharmaceutical Research, 1(2).

 doi:10.20944/preprints202603.0225.v1

7. Kumar, R., 2026. Process Optimization in Industrial Pharmaceutical Manufacturing Through Quality By Design (Qbd). International Journal of Drug Analysis, Industrial Pharmacy and Pharmaceutical Research, 1(2).
8. Fitzgerald, L., Niarchou, E., Jones, I. and Naughton, B., 2026. Emerging Digital Innovations in Pharmaceutical Manufacturing Quality: A Systematised Review. Journal of Pharmaceutical Innovation, 21(1), p.46.
9. Koo, J., Jeon, H., Cheong, J., Joo, Y., Han, S., Kim, H., Oh, J., Lee, D. and Oh, K.T., 2026. Modern approaches to quality by design for amorphous solid dispersion product development: a narrative review. Journal of Pharmaceutical Investigation, pp.1-19.
10. Gioumouxouzis, C.I., Eleftheriadis, G.K., Kyriakidis, A.S. and Karavasili, C., 2026. Translation of pharmaceutical 3D printing to clinical point-of-care and industrial manufacturing. Drug Delivery and Translational Research, pp.1-14.
11. Gadewar, S., Sangave, P., Saoji, S., Matte, K.V., Guntupalli, C., Hatware, K. and Pawde, D.M., 2026. Quality by Design (QbD)-Improved Polymeric Micelles of Posaconazole with Increased Solubility, Oral Bioavailability, and Antifungal Activity against Invasive Fungal Infections. Journal of Pharmaceutical Innovation, 21(2), p.130.
12. Karpicarov, D., Mitrevska, I., Manchevska, B., Apostolova, P., Tonic Ribarska, J. and Gjorgjeska, B., 2026. Software-assisted analytical Quality by Design for stability-indicating method development: integration of DoE and predictive retention modeling using MODDE® and DryLab®. Macedonian pharmaceutical bulletin, 72(2), pp.3-15.
13. Naeem, M.S., Naeem, M., Mehmood, M., Sulman, A.A. and Kharl, H.A.A., 2026. MACHINE LEARNING IN FORMULATION OPTIMIZATION AND PROCESS CONTROL. PharmA AI: Revolutionizing Drug Discovery and Healthcare with Artificial Intelligence.
14. Pérez-Beltrán, C.H., Jiménez-Carvelo, A.M., Sandoval-Sicairos, E.S., OSUNA-MARTÍNEZ, L.U., Santos-Ballardo, C.L., Carrazco-Ávila, P.Y., Cuevas-Rodríguez, E.O. and Cuadros-Rodríguez, L., 2026. Machine learning and optical imaging for pharmaceutical forensic toxicology: A comprehensive review. Journal of Pharmaceutical and Biomedical Analysis Open, p.100104.
15. Mariya, T., Parameswaran, K.M., Johny, A.P., Kamarajan, K. and Kunnambath, K., 2026. Machine Learning Assisted Simultaneous Estimation of Drugs in Multicomponent Formulations by Spectrophotometry. Indian Journal of Pharmaceutical Education & Research, 60.