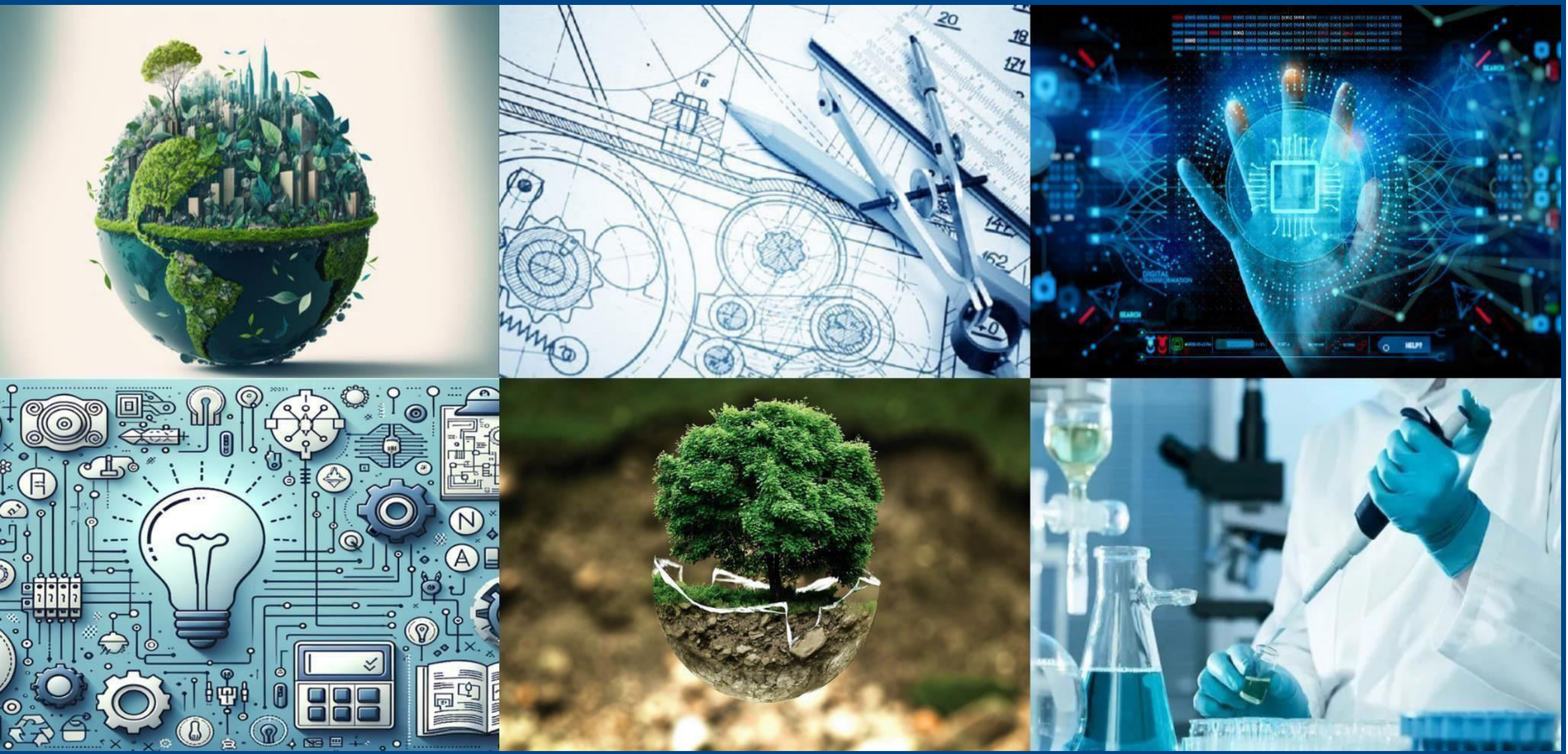
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206

Volume 8, Issue 9, September 2025

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Big Data-Driven Predictive Modeling of Pharmaceutical Tablet Quality using Machine Learning

Shaik Alima Zabi

Final Year MCA Student, School of Science and Computer Studies, CMR University, Bengaluru, Karnataka, India

ABSTRACT: This study rigorously examines the integration of machine learning (ML) methodologies to predict critical quality attributes (CQAs) in pharmaceutical tablet production. The research leverages a comprehensive, publicly available dataset encompassing 1,005 distinct manufacturing batches, each characterized by intricate raw material properties, varied process parameters, and an array of product quality outcomes. Six regression algorithms were deployed and systematically evaluated for their predictive performance. Among these, the Random Forest regressor demonstrated consistently superior results, achieving an R^2 value of 0.89 for impurity L —a notable benchmark in this context. In addition to traditional regression modeling, advanced feature selection and interpretability techniques such as Recursive Feature Elimination (RFE) and SHAP (SHapley Additive exPlanations) were employed. These methods facilitated the identification of influential variables, enhancing both real -time quality assurance and process optimization strategies in pharmaceutical manufacturing.

KEYWORDS: Pharmaceutical manufacturing, Critical Quality Attributes (CQAs), Predictive modeling, Random Forest, Machine learning

I. INTRODUCTION

Pharmaceutical manufacturing is a highly regulated and multifaceted discipline, wherein numerous process variables —including compression force, fill depth, and raw material attributes —collectively determine product quality, patient safety, and compliance w ith stringent regulatory standards. Historically, the industry has relied on end -product testing as the principal mechanism for quality assurance. While effective to some extent, this retrospective approach is inherently reactive, often resulting in delayed detection of quality deviations, increased production waste, and substantial financial implications.

With the advent of big data analytics and sophisticated machine learning techniques, the pharmaceutical sector is experiencing a paradigm shift toward proactive quality management. Predictive modeling, empowered by large -scale datasets, enables manufacturers to anticipate CQA outcomes such as impurity content, drug release kinetics, and residual solvent concentrations. This capability not only minimizes batch -to-batch variability but also strengthens product consistency and regulatory compliance.

The present research is focused on the application of ML -based predictive modeling to a cholesterol -lowering drug dataset. The objectives are threefold: (1) to develop accurate predictive models for six CQAs; (2) to systematically compare the predictive performance of several regression algorithms; and (3) to identify and interpret key process parameters influencing CQA outcomes through advanced feature selection and interpretability tools.

II. LITERATURE SURVEY

Recent advancements in pharmaceutical manufacturing underscore the importance of integrating Process Analytical Technology (PAT) and Quality by Design (QbD) frameworks. These paradigms advocate for data -driven, predictive approaches to process control and quality assurance. For instance, Zagar & Mihelic (2021) curated extensive datasets that have proven invaluable for ML applications in manufacturing variability analysis. Nasiri et al. (2019) illustrated that ensemble learning models can predict drug release characteristics with greater accuracy than traditional approaches. Rantanen & Khinast (2015) emphasized the synergy between PAT sensors and predictive analytics in delivering robust, real -time quality assurance. Additionally, Yu (2008) highlighted the FD A's QbD initiative, which encourages the adoption of predictive control strategies to enhance product robustness and regulatory compliance.

Despite these advances, there remains a notable gap in the literature regarding the systematic application of cutting -edge feature selection (such as Recursive Feature

Elimination) and interpretability methods (including SHAP) to pharmaceutical tablet quality prediction. This research addresses this deficiency by not only benchmarking predictive models, but also offering deeper insights into the underlying process variables that most significantly influence quality outcomes.

III. METHODOLOGY / APPROACH

Dataset

The "Cholesterol -Lowering Drug Process and Quality Data" set comprises 1,005 distinct manufacturing batches. Each sample is characterized by a rich set of variables, including:

- Raw material properties (e.g., excipient characteristics, moisture content)
- Process parameters (e.g., compression force, fill depth, press speed, startup dynamics)
- Quality outcomes (e.g., total impurities, average and minimum drug release percentage, residual solvent, impurity O, impurity L)

Preprocessing

Data preprocessing was conducted to ensure the integrity and suitability of the dataset for ML modeling. Missing numerical values were imputed using the median, while missing categorical values were replaced with the mode. Categorical variables underwent one-hot encoding to facilitate their inclusion in regression models. Numerical features were standardized to mitigate the effects of scale disparities among predictors.

Modeling

The study evaluated six regression algorithms:

- Linear Regression, Ridge, and Lasso (serving as linear baselines)
- Decision Tree (to capture nonlinear relationships)
- Random Forest (an ensemble method known for robustness)
- Support Vector Regression (utilizing kernel -based modeling)

The dataset was partitioned into a 70/30 train -test split, and model evaluation was further refined through 5 -fold cross -validation. Performance metrics included the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE), providing a comprehensive assessment of predictive accuracy.

Feature Selection & Interpretability

To enhance model interpretability and reduce dimensionality, Recursive Feature Elimination (RFE) was employed to identify the ten most predictive variables for each CQA. SHAP analysis was subsequently applied to elucidate the relative importance of these features, enabling the exclusion of variables with negligible impact (e.g., startup waste, initial press duration). This dual approach not only optimized model performance but also provided actionable insights into process optimization.

IV. RESULTS & DISCUSSION

Model Performance

Table 1 compares model performance across CQAs. Random Forest consistently outperformed other regressors, achieving $R^2 = 0.81$ for total impurities and $R^2 = 0.89$ for impurity L, with low error values. Linear models explained limited variance, while Decision Trees showed overfitting tendencies.

Table 1. Model performance for CQAs (best-performing model shown)

Target Variable	Best Model	R²	MAE	RMSE
Total Impurities	Random Forest	0.81	0.0228	0.0441
Drug Release Avg (%)	Random Forest	0.36	2.0600	2.6938
Drug Release Min (%)	Random Forest	0.30	2.6962	3.4965
Residual Solvent	Random Forest	0.61	0.0172	0.0265
Impurity O	Random Forest	0.05	0.0041	0.0086
Impurity L	Random Forest	0.87	0.0067	0.0114

The empirical findings underscore the effectiveness of ensemble learning —specifically the Random Forest regressor — in pharmaceutical tablet quality prediction. The Random Forest model consistently outperformed alternative algorithms, achieving an R² of 0.81 for total impurities and 0.89 for impurity L. These results suggest a high degree of predictive fidelity, particularly in the context of complex, multivariate manufacturing processes. Notably, the superior performance of Random Forest is attributable to its ability to capture intricate nonlinear relationships and interactions among variables, which are often present in pharmaceutical manufacturing data.

Beyond predictive accuracy, the application of RFE and SHAP facilitated a more nuanced understanding of the process variables most influential to CQA outcomes. For instance, factors such as compression force and moisture content were consistently identified as critical determinants of impurity levels and drug release profiles. By enabling real -time identification of key drivers of quality variation, these techniques support the implementation of proactive process controls and continuous improvement initiatives.

In summary, this research demonstrates that ML -driven predictive modeling, underpinned by robust feature selection and interpretability frameworks, holds significant promise for advancing pharmaceutical manufacturing quality assurance. The integration of these methods can lead to more efficient, reliable, and compliant production processes, ultimately benefiting both manufacturers and patients.

Feature Importance

In the context of process parameter significance, several variables clearly distinguish themselves as pivotal contributors to model performance. Notably, fill depth mean, compression force variability, cylinder height mean, and ejection mean emerge as the principal predictors influencing the targeted quality attributes. Their strong impact is consistent with established process understanding, where material distribution, force application, and mechanical movement directly affect tablet characteristics. On the other hand, variables such as the weekend indicator, total waste, and startup waste reveal minimal predictive value in this setting, indicating that operational timing or aggregate waste metrics may not capture the nuanced process dynamics that drive critical quality outcomes.

V. DISCUSSION

The robust prediction of total impurities and impurity L underscores Random Forest's capacity to handle nonlinear, multivariate relationships inherent in pharmaceutical manufacturing. This finding aligns with the broader literature, which recognizes ensemble methods as particularly adept at modeling complex interactions between process variables and quality responses. The model's moderate accuracy for drug release and residual solvent, however, suggests there are limitations to the information captured solely by current process and sensor data. This points to the likely value of incorporating additional features —such as detailed formulation properties or real -time environmental measurements — to further explain variance and enhance predictive accuracy.

In contrast, the model's inability to reliably predict impurity O highlights a significant gap in the present sensing infrastructure. This deficiency may stem from unmeasured factors or subtle process mechanisms not adequately represented by the available data streams. Addressing this gap will require both expanded sensor coverage and deeper process understanding.

VI. CONCLUSION

The evidence presented here demonstrates that integrating Random Forest regression with Recursive Feature Elimination (RFE) and SHAP (SHapley Additive exPlanations) provides a powerful framework for predicting critical quality attributes (CQAs) in tablet manufacturing. By systematically identifying and quantifying the influence of key process parameters, this approach enables more precise and responsive process control. Such capabilities are essential for advancing real -time quality monitoring, facilitating proactive process adjustments, and minimizing the production of off -specification product.

Looking ahead, the integration of advanced sensor technologies —such as near -infrared (NIR) spectroscopy —holds significant promise for further improving model fidelity, especially for attributes like impurity O that currently elude accurate prediction. Additionally, the exploration of closed -loop control strategies, where predictive insights inform automated process adjustments, represents a logical and impactful next step. This direction is supported by ongoing trends toward digitalization and automation in pharmaceutical manufacturing and will be instrumental in achieving consistently high product quality in increasingly complex manufacturing environments.

REFERENCES

[1] Zagar, J., & Mihelic, J. (2021). Big Data collection in pharmaceutical manufacturing. Figshare. https://doi.org/10.6084/m9.figshare.c.5645578.v3
[2] Yu, L. X. (2008). Pharmaceutical quality by design: Product and process development, understanding, and control. Pharmaceutical Research, 25(4), 781 –791. https://doi.org/10.1007/s11095 -007 -9511 -1
[3] Rantanen, J., & Khinast, J. (2015). The future of pharmaceutical manufacturing sciences. Journal of Pharmaceutical Sciences, 104(11), 3612 –3638. https://doi.org/10.1002/jps.24594
[4] Nasiri, H., Ghahremani, P., & Jahanmiri, A. (2019). Predictive modeling in pharmaceutical manufacturing using machine learning approaches. International Journal of Pharmaceutics, 567, 118445. https://doi.org/10.1016/j.ijpharm.2019.118445
[5] FDA. (2004). Guidance for Industry: PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance. U.S. Food and Drug Administration. https://www.fda.gov/media/71012/download
[6] ICH. (2009). ICH Harmonised Tripartite Guideline Q8(R2): Pharmaceutical Development. International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. https://www.ich.org/page/quality -guidelines
[7] Goh, G. B., Hodas, N. O., & Vishnu, A. (2017). Deep learning for computational chemistry. Journal of Computational Chemistry, 38(16), 1291 –1307. https://doi.org/10.1002/jcc.24764
[8] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com