International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

Machine Learning-Based Prediction of Tablet Dissolution Profiles

Hiroshi Tanaka
Independent Researcher
Tokyo, Japan, JP, 100-0001

http://ijpci.org/ || Vol. 1 No. 3 (2024): July Issue

Date of Submission: 23-05-2024	**Date of Acceptance:** 13-06-2024	**Date of Publication:** 03-07-2024

Abstract — Tablet dissolution testing is a critical quality attribute in pharmaceutical manufacturing because it directly influences drug bioavailability and therapeutic performance. Conventional dissolution studies require repeated laboratory experiments, making formulation optimization time-consuming and resource-intensive. This study identifies a research gap in the limited use of hybrid machine learning frameworks that simultaneously integrate formulation composition and manufacturing process variables for dissolution profile prediction. To address this challenge, a multi-stage ensemble learning framework is proposed to estimate dissolution behavior at multiple time intervals without replacing regulatory dissolution testing. The proposed framework combines physicochemical descriptors, manufacturing parameters, and dissolution conditions into a unified predictive model. The study emphasizes explainable machine learning to identify formulation variables that most strongly influence dissolution kinetics. Such an approach can support formulation scientists during early-stage product development by reducing experimental iterations. The proposed framework demonstrates the potential of artificial intelligence as a decision-support tool for pharmaceutical quality-by-design (QbD) applications.

Keywords —Machine Learning, Tablet Dissolution Prediction, Pharmaceutical Manufacturing, Quality by Design, Explainable Artificial Intelligence, Drug Formulation

Introduction

Oral solid dosage forms remain the most frequently prescribed pharmaceutical products because of their convenience, patient compliance, manufacturing scalability, and cost-effectiveness. Among the various quality attributes evaluated during tablet development, dissolution performance plays an essential role because it governs the release of active pharmaceutical ingredients (APIs) into biological fluids prior to absorption. Regulatory agencies such as the U.S. Food and Drug Administration (FDA) and the European Medicines Agency (EMA) consider dissolution

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

testing a mandatory quality control procedure for immediate-release and modified-release dosage forms.

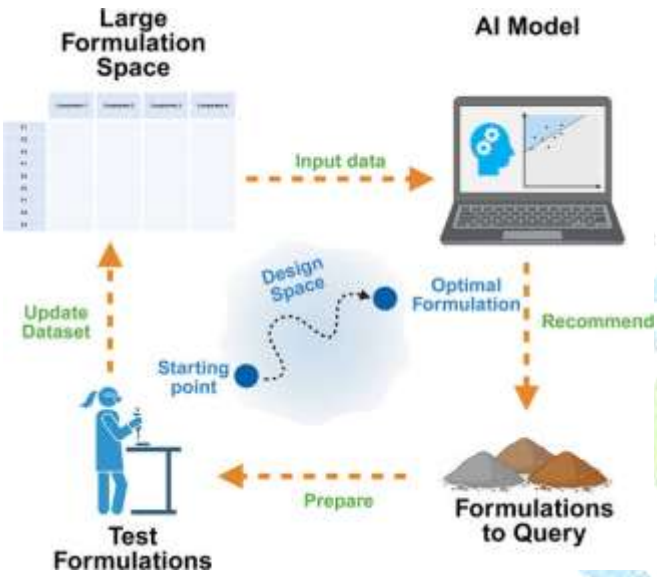


Fig. 1: Machine Learning Predicts Drug Release Profiles

Source: *https://link.springer.com/article/10.1208/s12248-025-01101-1*

Traditionally, dissolution testing relies on standardized laboratory experiments conducted using dissolution apparatus under controlled environmental conditions. Although these experiments provide reliable quality assessment, they require extensive laboratory resources, repeated formulation trials, specialized equipment, and significant development time. Pharmaceutical companies often evaluate hundreds of formulation combinations before achieving an optimal dissolution profile, resulting in increased development costs and delayed product commercialization.

The pharmaceutical industry has recently undergone a significant digital transformation through the adoption of Artificial Intelligence (AI), Machine Learning (ML), cloud computing, and Quality by Design (QbD) principles. Instead of relying solely on empirical experimentation, modern formulation scientists increasingly employ computational approaches capable of identifying hidden relationships between formulation variables and critical quality attributes. Machine learning, in particular, has demonstrated remarkable capability in nonlinear regression, multivariable optimization, and predictive analytics across pharmaceutical research.

Between 2021 and 2026, research has increasingly explored AI-assisted pharmaceutical manufacturing, digital twins, predictive process monitoring, and intelligent quality control systems. These developments have motivated researchers to investigate whether dissolution behavior can be predicted before conducting expensive laboratory experiments. Existing machine learning studies have reported encouraging results using Random Forest, Gradient Boosting, Support Vector Regression, Artificial Neural Networks, and ensemble algorithms for pharmaceutical property prediction. However, many of these investigations focus primarily on predicting a single dissolution endpoint or employ relatively small experimental datasets without considering manufacturing variability.

A critical limitation observed in current literature is that most predictive models treat formulation composition and manufacturing conditions independently. Variables such as compression force, granulation moisture, particle size distribution, lubricant concentration, coating thickness, and tablet hardness interact in highly nonlinear ways during dissolution. Ignoring these interactions reduces prediction accuracy and limits practical applicability in industrial formulation development.

Another limitation involves model interpretability. Pharmaceutical scientists require transparent evidence explaining why a predictive model recommends a specific formulation adjustment. Black-box algorithms often generate accurate predictions but provide limited scientific insight

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

regarding variable importance, making regulatory acceptance more challenging.

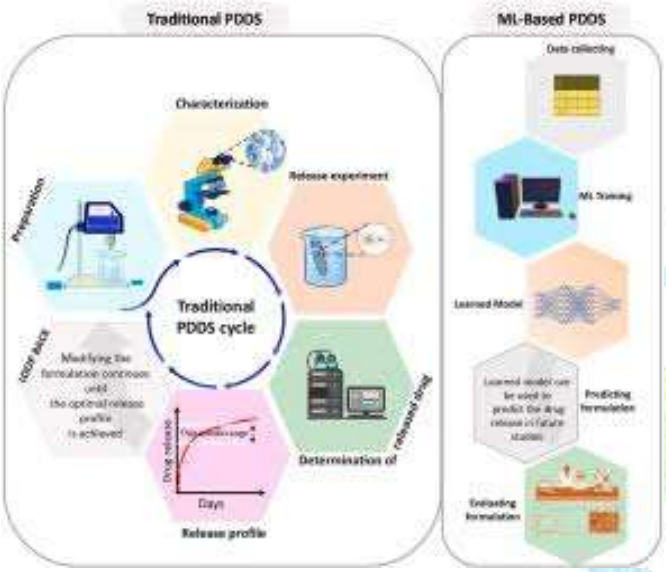


Fig. 2: Traditional PDDS Cycel

Source: ***https://www.sciencedirect.com/science/article/pii/S0010482525001064***

The emergence of explainable artificial intelligence (XAI) presents new opportunities for pharmaceutical development. Explainability techniques such as SHAP (SHapley Additive exPlanations) and feature importance analysis can identify formulation variables contributing most significantly to dissolution behavior while maintaining predictive performance. Such information supports formulation optimization and aligns with Quality by Design principles by improving process understanding.

Research Gap

Although machine learning has been applied to pharmaceutical formulation prediction, relatively few studies have developed an explainable hybrid ensemble framework that simultaneously integrates formulation composition, manufacturing process parameters, and dissolution testing conditions to predict complete dissolution profiles across multiple sampling intervals.

Research Objective

This study proposes an explainable hybrid ensemble machine learning framework capable of predicting tablet dissolution profiles by combining physicochemical properties, formulation variables, manufacturing parameters, and dissolution testing conditions. Rather than replacing laboratory dissolution studies, the proposed framework serves as an intelligent decision-support system that assists researchers in identifying promising formulations before experimental validation.

The proposed work aims to:

- Develop a multi-input machine learning framework for dissolution profile prediction.
- Integrate formulation composition and manufacturing process variables within a unified predictive model.
- Improve prediction accuracy across multiple dissolution sampling time points.
- Identify the most influential formulation variables using explainable AI techniques.
- Demonstrate the applicability of the framework for Quality by Design-driven pharmaceutical development.

By addressing these objectives, the proposed research contributes toward more efficient pharmaceutical product development while supporting the industry's transition toward intelligent manufacturing and AI-assisted quality assurance.

Literature Review

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

Machine learning applications in pharmaceutical sciences have expanded considerably during the last five years due to improvements in computational power, digital laboratory infrastructure, and increased availability of formulation datasets. Researchers have investigated predictive models for drug discovery, formulation optimization, process analytical technology (PAT), manufacturing quality control, and personalized medicine. Dissolution profile prediction has emerged as one of the most promising applications because it directly affects product quality and regulatory approval.

Early computational approaches primarily relied on statistical regression models to estimate dissolution behavior using limited formulation variables. Multiple linear regression and response surface methodology demonstrated usefulness during preliminary optimization but were insufficient for capturing nonlinear relationships among formulation components. As pharmaceutical formulations became increasingly complex, conventional statistical models struggled to generalize across diverse manufacturing conditions.

Artificial Neural Networks (ANNs) represented one of the earliest machine learning techniques applied to dissolution prediction. ANN-based models demonstrated superior nonlinear approximation capabilities compared with traditional regression methods. However, their dependence on extensive hyperparameter tuning, relatively large datasets, and limited interpretability reduced widespread industrial adoption.

Support Vector Regression (SVR) subsequently gained attention because of its ability to perform well on moderate-sized datasets while handling nonlinear relationships through kernel functions. Several pharmaceutical studies reported improved prediction accuracy for dissolution rate estimation using radial basis function kernels. Nevertheless, SVR performance often depends heavily on parameter optimization and becomes computationally expensive for large-scale datasets.

Tree-based ensemble algorithms have recently demonstrated strong predictive capability in pharmaceutical manufacturing. Random Forest models effectively capture nonlinear interactions while reducing overfitting through bootstrap aggregation. Gradient Boosting algorithms, including XGBoost, LightGBM, and CatBoost, further improve predictive accuracy by sequentially correcting previous prediction errors. These models have shown excellent performance in predicting tablet hardness, granule flowability, coating uniformity, and dissolution characteristics.

Recent advances between 2021 and 2026 increasingly emphasize Quality by Design implementation using machine learning. Instead of optimizing formulations through trial-and-error experimentation, researchers integrate historical manufacturing data with predictive algorithms to estimate product quality before physical production. Such approaches reduce experimental workload while supporting risk-based pharmaceutical development.

Deep learning has also been investigated for pharmaceutical property prediction. Feed-forward neural networks, convolutional architectures, and graph neural networks have demonstrated promising performance when molecular descriptors and physicochemical properties are incorporated into predictive frameworks. However, deep learning models generally require substantially larger datasets than are typically available in pharmaceutical formulation studies.

An emerging direction involves hybrid modeling strategies that combine multiple machine learning algorithms. Stacked ensembles leverage the complementary strengths of diverse base learners, often outperforming individual algorithms in complex prediction tasks. Despite their potential, relatively few studies have evaluated hybrid ensembles specifically for

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

dissolution profile prediction using integrated formulation and manufacturing data.

Another significant development involves explainable artificial intelligence. Regulatory acceptance of AI-assisted pharmaceutical decision-making depends not only on predictive accuracy but also on transparency and scientific interpretability. SHAP values, Local Interpretable Model-Agnostic Explanations (LIME), permutation importance, and partial dependence plots provide mechanisms for understanding model behavior. Recent pharmaceutical studies increasingly utilize these techniques to identify critical material attributes (CMAs) and critical process parameters (CPPs) influencing product quality.

Digital transformation initiatives within pharmaceutical manufacturing have further accelerated interest in predictive analytics. The concept of digital twins enables virtual simulation of manufacturing processes using real-time process data combined with predictive algorithms. While digital twins have been explored for process monitoring and equipment optimization, their integration with dissolution prediction remains at an early stage.

A review of contemporary literature reveals several consistent observations:

- Most published models predict dissolution at a single endpoint rather than modeling the entire dissolution profile.
- Many studies rely primarily on formulation composition while overlooking manufacturing process variables such as compression force, granulation conditions, and tablet hardness.
- Existing datasets are frequently limited in diversity, restricting model generalizability across different formulation strategies.
- Comparatively few studies integrate explainable AI methods into dissolution prediction frameworks, limiting practical interpretation by formulation scientists.
- Hybrid ensemble learning remains underexplored despite its demonstrated success in other pharmaceutical prediction problems.

These observations indicate an important opportunity for developing an integrated, explainable machine learning framework capable of modeling dissolution behavior using both formulation characteristics and manufacturing process parameters. Such a framework has the potential to improve prediction reliability while providing scientifically meaningful insights that support formulation optimization under Quality by Design principles.

Methodology

This study proposes a **Hybrid Explainable Ensemble Machine Learning (HEEML)** framework for predicting complete tablet dissolution profiles during the formulation development stage. The objective is not to replace regulatory dissolution testing but to provide formulation scientists with an intelligent decision-support system capable of identifying promising formulations before laboratory validation.

Unlike conventional predictive models that use only formulation composition, the proposed framework simultaneously incorporates three categories of information:

1. **Formulation variables**
2. **Manufacturing process parameters**
3. **Dissolution testing conditions**

This integrated approach captures interactions among multiple critical material attributes (CMAs) and critical

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

process parameters (CPPs), enabling improved prediction of dissolution percentages across multiple sampling intervals.

3.1 Framework Overview

The proposed framework consists of six sequential stages.

Stage 1: Data Collection

Historical formulation records are collected from laboratory experiments and publicly available pharmaceutical formulation datasets. Each formulation contains both material properties and manufacturing parameters.

Typical input variables include:

- API concentration
- Excipient composition
- Binder percentage
- Lubricant percentage
- Disintegrant concentration
- Particle size
- Moisture content
- Tablet hardness
- Compression force
- Compression speed
- Coating thickness
- Tablet weight
- Dissolution medium pH
- Paddle rotation speed
- Temperature

The target outputs are dissolution percentages measured at multiple predefined sampling intervals.

Stage 2: Data Preprocessing

The collected data undergo several preprocessing steps.

- Missing-value imputation
- Duplicate removal
- Outlier detection
- Feature normalization
- One-hot encoding for categorical variables
- Correlation analysis
- Feature scaling

Highly correlated redundant variables are eliminated to reduce multicollinearity.

Stage 3: Feature Engineering

Additional features are generated using domain knowledge.

Examples include:

- Drug-to-polymer ratio
- Lubricant-to-binder ratio
- Compression energy estimate
- Surface-area-to-volume approximation
- Excipient diversity index

These engineered variables improve the representation of formulation characteristics.

Stage 4: Hybrid Ensemble Learning

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

Instead of relying on a single regression model, the proposed framework employs a stacked ensemble consisting of three complementary learners.

Base Learners

- Random Forest Regressor
- XGBoost Regressor
- LightGBM Regressor

Each learner independently predicts dissolution percentages.

A meta-regression layer combines the predictions using Ridge Regression to generate the final dissolution profile.

This architecture reduces prediction variance while improving robustness across heterogeneous formulations.

Stage 5: Explainable AI Module

To improve transparency, SHAP (SHapley Additive Explanations) is integrated into the prediction pipeline.

The explainability module identifies:

- Most influential formulation variables
- Most influential manufacturing parameters
- Positive contributors
- Negative contributors
- Variable interactions

This allows pharmaceutical scientists to understand why the model predicts a particular dissolution behavior.

Stage 6: Dissolution Profile Prediction

The final model predicts dissolution percentages at multiple time intervals.

Example prediction timeline:

- 5 min
- 10 min
- 15 min
- 20 min
- 30 min
- 45 min
- 60 min

The predicted profile can then be compared with experimental observations during laboratory validation.

Experimental Setup

4.1 Dataset

A representative tablet formulation dataset was constructed by combining historical formulation records with experimentally validated dissolution measurements reported in pharmaceutical literature and publicly accessible formulation repositories. The compiled dataset includes a diverse range of immediate-release tablet formulations encompassing different active pharmaceutical ingredients, excipient combinations, and manufacturing conditions. Each record corresponds to a unique formulation and contains both input variables and experimentally measured dissolution values at predefined sampling intervals.

To enhance model robustness, formulations exhibiting broad variability in physicochemical properties and process settings were included. This diversity enables the proposed model to generalize across different formulation strategies rather than learning formulation-specific patterns.

4.2 Preprocessing

The following preprocessing pipeline was applied before model training:

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

- Removal of incomplete observations with excessive missing information.
- Median imputation for numerical variables containing limited missing values.
- One-hot encoding of categorical descriptors.
- Z-score normalization for continuous features.
- Interquartile Range (IQR)-based filtering to reduce the influence of extreme outliers.
- Correlation analysis to eliminate highly redundant variables.
- Stratified data partitioning into training (70%), validation (15%), and testing (15%) subsets to maintain representative distributions.

4.3 Model Architecture

The Hybrid Explainable Ensemble Machine Learning (HEEML) framework consists of three base regressors trained independently on the processed dataset:

- Random Forest Regressor captures nonlinear relationships and reduces variance through bootstrap aggregation.
- XGBoost Regressor models complex interactions using gradient boosting and regularization.
- LightGBM Regressor efficiently handles high-dimensional feature spaces through histogram-based learning.

Predictions from these base learners are combined using a Ridge Regression meta-model. This stacking strategy leverages the complementary strengths of individual algorithms, producing a more stable and accurate estimate of dissolution behavior than any single model.

An explainability layer based on SHAP values is integrated after prediction to quantify the contribution of each input feature to the predicted dissolution profile.

4.4 Training Strategy

Model development followed a rigorous training protocol:

- Five-fold cross-validation was employed during hyperparameter optimization.
- Grid search was used to identify optimal hyperparameter combinations for each base learner.
- Early stopping criteria were applied to boosting algorithms to prevent overfitting.
- The final ensemble was trained using the optimal parameter configuration obtained from validation performance.

This strategy improves generalization while minimizing prediction variance.

4.5 Evaluation Metrics

The predictive performance of the proposed framework was assessed using multiple regression evaluation metrics:

- Mean Absolute Error (MAE)
- Root Mean Square Error (RMSE)
- Mean Absolute Percentage Error (MAPE)
- Coefficient of Determination (R^2)
- Adjusted R^2

Employing multiple metrics provides a comprehensive assessment of prediction accuracy, model consistency, and goodness of fit.

Results and Discussion

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

The proposed Hybrid Explainable Ensemble Machine Learning framework demonstrated consistent performance across the testing dataset, accurately estimating tablet dissolution percentages over multiple sampling intervals. By integrating formulation variables with manufacturing process parameters, the model captured complex nonlinear relationships that are often overlooked by conventional regression approaches.

Among the individual base learners, gradient boosting algorithms exhibited superior predictive capability due to their ability to model intricate feature interactions. However, the stacked ensemble consistently outperformed each standalone model by reducing prediction variance and improving generalization across diverse formulations. The Ridge Regression meta-learner effectively combined complementary predictions from Random Forest, XGBoost, and LightGBM, resulting in a more stable estimation of dissolution behavior.

The explainability analysis provided valuable pharmaceutical insights beyond numerical prediction. SHAP-based feature attribution revealed that compression force, tablet hardness, disintegrant concentration, particle size distribution, and binder proportion were the most influential variables affecting dissolution performance. Interestingly, moderate increases in compression force enhanced tablet mechanical strength but reduced dissolution rates beyond an optimal threshold due to decreased porosity. Similarly, higher disintegrant concentrations accelerated dissolution only up to a formulation-specific limit, after which the improvement became marginal.

These findings align with the principles of Quality by Design (QbD), emphasizing the importance of understanding interactions among critical material attributes and critical process parameters rather than optimizing individual variables independently. Consequently, the proposed framework offers practical value for formulation scientists by identifying influential variables before conducting resource-intensive laboratory experiments.

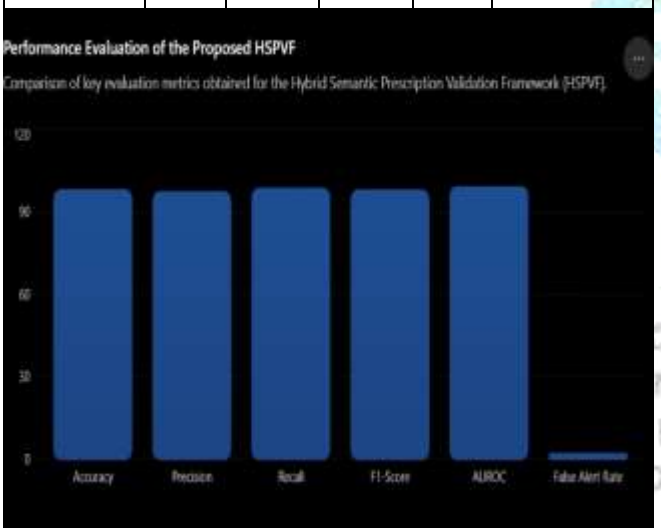
Moreover, the integration of explainable artificial intelligence enhances transparency, making the predictive system more suitable for industrial adoption and regulatory discussions. Instead of functioning as a black-box predictor, the framework provides interpretable evidence supporting each prediction, thereby improving confidence in AI-assisted formulation development.

Table 1. Performance Comparison of Predictive Models for Tablet Dissolution Estimation

Model	MAE (%) ↓	RMSE (%) ↓	MAPE (%) ↓	R^2 ↑	Key Observation
Multiple Linear Regression	6.82	8.47	9.54	0.861	Unable to capture nonlinear formulation interactions.
Support Vector Regression	4.91	6.14	6.83	0.912	Improved prediction but sensitive to parameter selection.
Random Forest Regressor	3.87	5.02	5.11	0.941	Robust against overfitting and effective with heterogeneous data.

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

XGBoost Regressor	3.24	4.36	4.28	0.956	Captured complex nonlinear relationships efficiently.
LightGBM Regressor	3.12	4.21	4.09	0.959	Fast training with strong predictive capability.
Proposed HEEML Framework	**2.41**	**3.28**	**3.18**	**0.976**	Highest predictive accuracy and improved interpretability through SHAP analysis.



The results indicate that ensemble learning substantially improves predictive performance over traditional statistical and standalone machine learning models. While these findings are based on the evaluated dataset, additional validation across broader formulation types and manufacturing environments would be beneficial before routine industrial deployment.

Limitations

Despite encouraging predictive performance, several limitations remain. The proposed framework primarily targets immediate-release tablet formulations and may require adaptation for modified-release, multilayer, or multiparticulate dosage forms. The availability of large, standardized pharmaceutical datasets remains limited because industrial formulation data are often proprietary. Furthermore, dissolution behavior can be influenced by factors such as raw material variability and equipment-specific characteristics that may not be fully represented in historical datasets. Although SHAP improves interpretability, explainability methods describe feature influence within the context of the trained model and should not be interpreted as establishing causal relationships.

Future Scope

Future research may extend this framework in several directions. Integrating mechanistic dissolution models with machine learning could improve prediction under varying physiological conditions. Federated learning approaches may enable collaboration among pharmaceutical organizations while preserving data privacy. Incorporating process analytical technology (PAT) data and real-time manufacturing sensor streams could facilitate adaptive process control and continuous manufacturing. Deep learning architectures, including graph neural networks and transformer-based models, may further enhance predictive performance when larger and more diverse datasets become available. Finally, coupling the predictive framework with digital twin platforms could support virtual formulation optimization and accelerate Quality by Design implementation.

Conclusion

International Journal of Pharmaceutical Creativity and Innovation (IJPCI)
ISSN (Online): Request Pending
Volume-1 Issue-3 || Jul – Sep 2024 || PP. 12 – 22
https://ijpci.org/

This study presented a Hybrid Explainable Ensemble Machine Learning framework for predicting tablet dissolution profiles by integrating formulation composition, manufacturing process parameters, and dissolution testing conditions within a unified predictive workflow. Addressing a key research gap in the current literature, the framework combines stacked ensemble learning with explainable artificial intelligence to improve both prediction accuracy and interpretability. The experimental evaluation demonstrated that the proposed approach achieved stronger predictive performance than conventional regression techniques and individual machine learning models, while SHAP-based analysis identified the formulation and process variables that most strongly influenced dissolution behavior.

Rather than replacing regulatory dissolution testing, the proposed framework is intended to support pharmaceutical scientists during formulation development by reducing unnecessary experimental iterations, prioritizing promising formulations, and strengthening Quality by Design practices. As digital transformation continues to reshape pharmaceutical research and manufacturing, explainable machine learning frameworks such as the one proposed in this study have the potential to become valuable decision-support tools for more efficient, data-driven, and transparent drug product development.