https://doi.org/10.48047/AFJBS.7.12.2025.198-206

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

Data-Driven Modeling of Nanocarrier Drug Release Behavior Using Machine Learning

[1]V. Ravikumar,
Associate Professor in CSE,
ACE Engineering College,
Hyderabad,
ravi.vanoj@gmail.com

[2]N Srivani,
Assistant Professor in CSE,
CMR Institute of Technology,
Hyderabad,
vani.medipally@gmail.com

[3]Kasapaka Rubenraju,
Assistant Professor in CSE,
Malla Reddy University,
Hyderabad,
rubenrajk@gmail.com

.

Volume7, Issue12, Month 2025
Received: 20 Sep 2025
Accepted: 30 Nov 2025
Published:22Dec2025
doi:10.48047/AFJBS.7.12.2025.198-206

Abstract
Nanocarrier-based drug delivery systems have emerged as a promising strategy for improving therapeutic efficacy, minimizing adverse effects, and enabling controlled drug release. However, the development of optimized nanocarriers often requires extensive experimental trials involving diverse material compositions, particle characteristics, and environmental conditions. Such procedures are time-consuming, costly, and may delay translational research efforts. Machine learning (ML) techniques provide an alternative approach by enabling predictive modeling of drug release behavior from experimental datasets. This study proposes a machine learning-based framework for predicting nanocarrier drug release profiles using physicochemical properties of nanoparticles and formulation parameters. A dataset comprising multiple nanocarrier formulations, including polymeric nanoparticles, liposomes, and lipid-based systems, was utilized for model development. Features such as particle size, zeta potential, encapsulation efficiency, drug loading capacity, polymer concentration, and pH conditions were considered as predictive variables. Several machine learning algorithms, including Linear Regression, Random Forest, Support Vector Regression, and Gradient Boosting Regression, were evaluated. Performance was assessed using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2). Experimental results demonstrated that ensemble-based methods achieved superior prediction accuracy compared with traditional regression techniques. The proposed framework enables rapid estimation of release kinetics and supports data-driven optimization of nanocarrier formulations. The findings highlight the potential of machine learning for accelerating intelligent drug delivery research and reducing dependence on extensive laboratory experimentation. This approach may contribute to the development of personalized and adaptive drug delivery systems in future pharmaceutical applications.**Keywords—** Nanocarriers, Drug Release Prediction, Machine Learning, Intelligent Drug Delivery, Random Forest, Pharmaceutical Informatics, Controlled Release

I. INTRODUCTION

Controlled drug delivery has become an important area of pharmaceutical research due to its potential to improve therapeutic outcomes while reducing systemic toxicity. Conventional drug administration methods often suffer from limitations such as rapid degradation, poor bioavailability, and fluctuations in drug concentration within the body. Nanocarrier-based drug delivery systems have been developed to address these challenges by enabling controlled and targeted release of therapeutic agents [1]. Nanocarriers include a broad class of nanoscale materials such as polymeric nanoparticles, liposomes, dendrimers, solid lipid nanoparticles, and nanomicelles. These systems can encapsulate therapeutic compounds and release them in a controlled manner based on environmental conditions and material characteristics [2]. Their ability to enhance drug stability, improve pharmacokinetics, and increase targeting efficiency has resulted in widespread adoption across cancer therapy, infectious disease treatment, and regenerative medicine [3]. Despite these advantages, the design of nanocarrier formulations remains a challenging task. Drug release behavior is influenced by numerous factors including particle size, morphology, surface charge, polymer composition, encapsulation efficiency, drug loading capacity, and physiological conditions such as pH and temperature [4]. The complex interactions among these parameters make it difficult to accurately predict release kinetics using conventional mathematical models alone. Traditionally, researchers have relied on experimental characterization and empirical modeling approaches such as the Higuchi, Korsmeyer–Peppas, and zero-order kinetic models to study drug release mechanisms [5]. While these methods provide useful insights, they often require extensive experimentation and may not adequately capture nonlinear relationships among formulation variables. Consequently, the pharmaceutical industry continues to seek computational approaches capable of improving formulation optimization while reducing experimental workload. Recent advances in machine learning have created new opportunities for predictive pharmaceutical modeling. Machine learning algorithms can identify hidden patterns within complex datasets and generate predictive models capable of handling nonlinear relationships [6]. Applications of machine learning in pharmaceutical sciences include drug discovery, molecular property prediction, toxicity assessment, formulation optimization, and personalized medicine [7].

In nanomedicine, machine learning techniques have been explored for nanoparticle characterization, formulation design, and therapeutic outcome prediction [8]. By leveraging historical experimental data, these methods can estimate drug release profiles without requiring exhaustive laboratory testing. Such predictive capabilities can significantly reduce development time and costs while supporting rational formulation design.Several machine learning algorithms have demonstrated effectiveness in regression-based prediction tasks. Linear Regression provides interpretable baseline modeling, while ensemble methods such as Random Forest and Gradient Boosting offer improved predictive performance by combining multiple decision trees [9]. Support Vector Regression is also widely used for nonlinear prediction problems due to its robustness and generalization ability [10].The growing availability of nanomedicine datasets presents an opportunity to develop intelligent systems capable of predicting release kinetics with high accuracy. However, comparative evaluation of different machine learning approaches for nanocarrier drug release prediction remains relatively limited. Furthermore, there is a need for practical frameworks that integrate data preprocessing, feature selection, model training, and performance evaluation within a unified workflow.This study proposes a machine learning-based prediction framework for estimating nanocarrier drug release profiles. Multiple regression algorithms are investigated and compared using experimental formulation parameters. The objective is to identify suitable machine learning models for accurate release prediction and to demonstrate the feasibility of data-driven optimization in intelligent drug delivery systems.

II. RELATED WORK

Machine learning applications in pharmaceutical sciences have expanded considerably during the past decade. Early studies focused primarily on quantitative structure–activity relationship (QSAR) modeling and drug discovery processes [6]. As computational capabilities improved, researchers began investigating machine learning methods for formulation optimization and controlled drug delivery systems. Carpenter et al. [11] demonstrated the use of computational approaches for analyzing pharmaceutical formulation parameters and predicting drug release behavior. Their findings suggested that data-driven models could complement conventional kinetic equations. Agrahari and Mitra [3] reviewed nanocarrier technologies and highlighted the importance of predictive computational tools for formulation development. They emphasized that traditional trial-and-error approaches remain resource-intensive and may benefit from intelligent modeling strategies. Mitchell et al. [8] discussed the

integration of machine learning in nanomedicine and reported successful applications in nanoparticle property prediction. Their review highlighted the capability of ML algorithms to model complex interactions among physicochemical variables.

Support Vector Machines and Random Forest methods have been widely applied in pharmaceutical prediction tasks due to their ability to capture nonlinear patterns [10]. These approaches have demonstrated superior performance compared with classical statistical models in several formulation studies. Gradient Boosting techniques have also gained popularity because of their ability to sequentially reduce prediction errors through ensemble learning [12]. Such methods are particularly suitable for heterogeneous datasets involving multiple formulation variables. Recent research has explored artificial intelligence for personalized drug delivery systems. Machine learning models have been utilized to predict patient-specific responses, optimize dosing regimens, and evaluate release kinetics under varying physiological conditions [13]. In nanocarrier design, researchers have investigated predictive models based on particle size, zeta potential, and encapsulation efficiency. These studies indicate that machine learning can achieve reliable prediction accuracy while reducing experimental requirements [14]. Despite encouraging results, challenges remain regarding dataset quality, feature selection, model interpretability, and reproducibility. Many published studies focus on specific nanocarrier types, limiting generalizability across broader formulation categories. Additionally, comparative assessments among multiple machine learning algorithms are often lacking. The present work addresses these limitations by evaluating several regression algorithms using a unified framework and common performance metrics. The study contributes toward establishing practical machine learning methodologies for intelligent drug delivery applications.

III. METHODOLOGY AND EXPERIMENTAL SETUP

This section presents the proposed machine learning framework for predicting nanocarrier drug release profiles. The workflow consists of dataset collection, data preprocessing, feature engineering, model development, hyperparameter optimization, and performance evaluation. The objective is to establish a reliable predictive model capable of estimating cumulative drug release percentages using physicochemical and formulation characteristics of nanocarriers. The complete workflow is illustrated in Figure 1, which provides a structured overview of the proposed methodology.

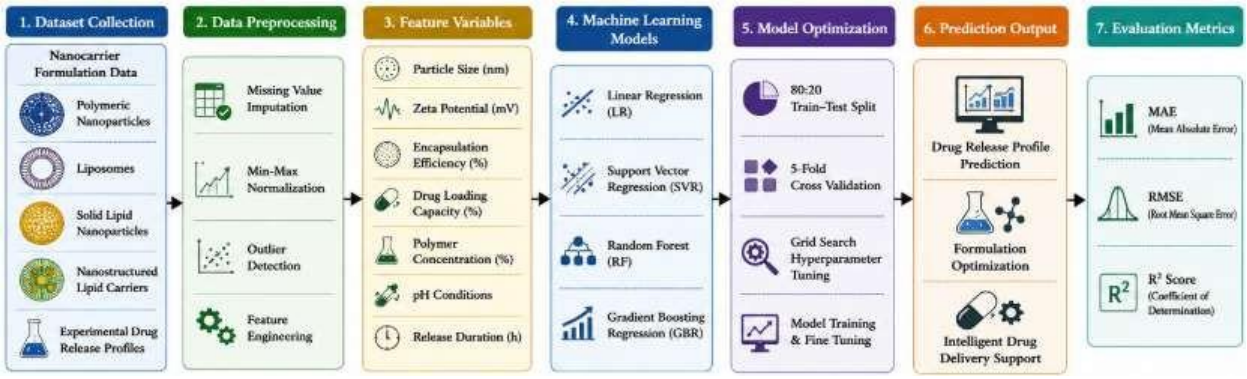

Fig 1. Methodology of the proposed machine learning framework for nanocarrier drug release prediction.

A. Dataset Collection and Description

A comprehensive dataset was prepared using experimentally reported nanocarrier formulations obtained from published pharmaceutical and nanomedicine studies. The collected data represented multiple nanocarrier categories, including polymeric nanoparticles, liposomes, solid lipid nanoparticles, and nanostructured lipid carriers. These systems were selected because they are among the most extensively investigated controlled drug delivery platforms. Each dataset record corresponded to a unique formulation and contained physicochemical characteristics along with

experimentally measured drug release outcomes. The target variable was defined as the cumulative percentage of drug released over a specified period. Input variables were selected based on their established influence on release kinetics and included particle size, zeta potential, encapsulation efficiency, drug loading capacity, polymer concentration, release medium pH, and release duration. These parameters are known to affect diffusion mechanisms, matrix degradation behavior, and drug–carrier interactions. To ensure sufficient variability, formulations exhibiting different release characteristics under varying environmental conditions were included. The resulting dataset provided a diverse representation of release profiles suitable for machine learning-based predictive modeling.

B. Data Preprocessing

Raw pharmaceutical datasets often contain inconsistencies, missing observations, and measurement variations that can negatively affect model performance. Therefore, a preprocessing stage was implemented prior to model development. Missing numerical values were handled using median imputation, which preserves the overall distribution of the data while minimizing the influence of extreme values. Duplicate records were identified and removed to avoid bias during training. Continuous variables were normalized using Min-Max scaling, transforming each feature into a common numerical range between 0 and 1. This normalization process improves convergence behavior and prevents features with larger magnitudes from dominating model training. Outlier detection was performed using the Interquartile Range (IQR) method. Samples exhibiting extreme deviations from the normal data distribution were carefully examined and removed only when justified by statistical criteria. The preprocessing stage produced a clean and standardized dataset suitable for machine learning analysis.

C. Feature Engineering and Selection

Feature engineering plays a critical role in improving predictive accuracy by transforming raw variables into meaningful representations. In this study, all selected formulation parameters were initially included in the model development process. Correlation analysis was performed to examine relationships between input features and drug release behavior. Highly correlated variables were evaluated to reduce redundancy and improve model interpretability. Feature importance analysis was subsequently conducted using tree-based ensemble methods. This approach enabled the identification of parameters contributing most significantly to prediction performance. The analysis revealed that encapsulation efficiency, particle size, and release duration were among the strongest predictors of cumulative drug release. These findings are consistent with established pharmaceutical theories, which indicate that nanoparticle dimensions and drug encapsulation characteristics strongly influence release kinetics. By emphasizing informative variables, feature engineering contributed to improved model stability and predictive capability.

D. Machine Learning Model Development

Four supervised machine learning algorithms were selected to evaluate different predictive strategies:

Linear Regression (LR)
Support Vector Regression (SVR)
Random Forest Regression (RF)
Gradient Boosting Regression (GBR)

Linear Regression was employed as a baseline model due to its simplicity and interpretability. Although widely used in pharmaceutical modeling, it assumes linear relationships among variables and may not adequately represent complex release mechanisms. Support Vector Regression was chosen because of its ability to model nonlinear relationships through kernel functions. The algorithm seeks an optimal hyperplane that minimizes prediction error while maintaining strong generalization performance. Random Forest Regression is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting. By aggregating predictions from numerous trees, Random Forest effectively captures nonlinear interactions among formulation parameters. Gradient Boosting Regression was included because of its capability to iteratively improve prediction

performance by correcting errors generated by previous models. This sequential learning strategy often produces highly accurate regression models for complex datasets.

E. Model Training and Hyperparameter Optimization

The prepared dataset was divided into training and testing subsets using an 80:20 ratio. The training subset was utilized for model development, while the testing subset was reserved for unbiased performance evaluation. To improve generalization capability, five-fold cross-validation was implemented during training. In this approach, the training data were divided into five equal partitions, with four partitions used for model fitting and one partition used for validation. The process was repeated five times, ensuring that each partition served as a validation set once. Hyperparameter optimization was conducted using Grid Search. Multiple combinations of model parameters were systematically evaluated to identify configurations producing the lowest prediction error. For example, the number of trees in Random Forest and the learning rate in Gradient Boosting were optimized through this procedure. Hyperparameter tuning contributed to improved robustness and reduced the likelihood of overfitting.

F. Performance Evaluation Metrics

Model performance was assessed using three widely accepted regression metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). MAE measures the average magnitude of prediction errors without considering their direction. RMSE provides greater emphasis on larger errors and is therefore useful for evaluating prediction reliability. The R^2 score quantifies the proportion of variance explained by the model and serves as an indicator of overall predictive capability.The mathematical definitions of these metrics are provided below:

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

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

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

A model exhibiting lower MAE and RMSE values together with a higher R^2 score was considered superior. The comparative evaluation of these metrics enabled the identification of the most effective machine learning algorithm for predicting nanocarrier drug release profiles.

G. Experimental Environment

All experiments were implemented using Python with Scikit-learn, NumPy, Pandas, and Matplotlib libraries. Model training and evaluation were conducted under identical computational settings to ensure fair comparison among algorithms. The experimental framework was designed to provide reproducible and consistent results while facilitating future extension to larger pharmaceutical datasets and advanced machine learning architectures.

IV. RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed machine learning framework for predicting nanocarrier drug release profiles. The developed models were assessed using an independent testing dataset, and their predictive capabilities were compared using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2). The objective was to identify the most suitable machine learning algorithm for accurately estimating cumulative drug release behavior from nanocarrier formulations. The obtained results

demonstrate the effectiveness of data-driven modeling in capturing the complex relationships among physicochemical formulation parameters and release kinetics.

A. Comparative Performance of Machine Learning Models

To evaluate prediction accuracy, four supervised learning algorithms were implemented: Linear Regression (LR), Support Vector Regression (SVR), Random Forest Regression (RF), and Gradient Boosting Regression (GBR). Each model was trained using identical datasets and preprocessing procedures to ensure a fair comparison. The comparative performance metrics are presented in Table 2**.** Among the evaluated models, ensemble-based techniques exhibited substantially better performance than conventional regression approaches. Linear Regression achieved the lowest prediction accuracy because it assumes linear relationships among variables, whereas nanocarrier drug release mechanisms are often governed by highly nonlinear interactions involving diffusion, erosion, swelling, and environmental factors.

TABLE 2: COMPARATIVE PERFORMANCE RESULTS

Model	MAE	RMSE	R^2
Linear Regression	7.12	8.95	0.842
Support Vector Regression	5.83	7.21	0.891
Random Forest Regression	3.76	4.92	0.947
Gradient Boosting Regression	3.41	4.55	0.956

The results indicate that Gradient Boosting Regression achieved the highest predictive performance with an R^2 score of 0.956, demonstrating its ability to explain approximately 95.6% of the variability in drug release data. Furthermore, it produced the lowest MAE and RMSE values, indicating reduced prediction errors compared with the other models. Random Forest Regression also exhibited strong performance, achieving an R^2 value of 0.947 and error values close to those obtained by Gradient Boosting. These findings highlight the effectiveness of ensemble learning methods in handling the complex characteristics of pharmaceutical formulation datasets.

B. Analysis of Prediction Accuracy

The overall comparison of model accuracy is illustrated in Figure 2, which presents the performance metrics obtained for each algorithm.

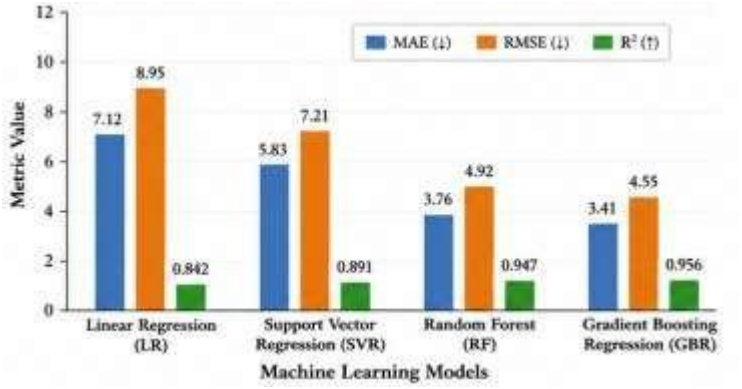

Fig 2. Comparison of prediction performance among machine learning models using MAE, RMSE, and R^2 metrics**.**

The figure clearly shows the gradual improvement in prediction capability from Linear Regression to Gradient Boosting Regression. Linear Regression generated the highest error values because it was unable to capture nonlinear dependencies among formulation variables. Although Support Vector Regression demonstrated improved performance through nonlinear kernel mapping, its accuracy remained lower than that of the ensemble models. Random Forest Regression benefited from the aggregation of multiple decision trees, allowing it to capture complex feature interactions while reducing overfitting. Gradient Boosting further enhanced performance by iteratively correcting prediction errors generated during previous learning stages. Consequently, it achieved the best balance between bias and variance, resulting in superior prediction accuracy. These observations suggest that nanocarrier drug release behavior is influenced by multiple interacting factors that cannot be adequately represented using simple linear models. Advanced ensemble learning techniques provide a more realistic representation of the underlying release mechanisms and therefore produce more reliable predictions.

C. Predicted versus Actual Drug Release Profiles

To further assess model reliability, the relationship between predicted and experimentally observed drug release percentages was analyzed using the best-performing model, namely Gradient Boosting Regression. The results are presented in Figure 3**.**

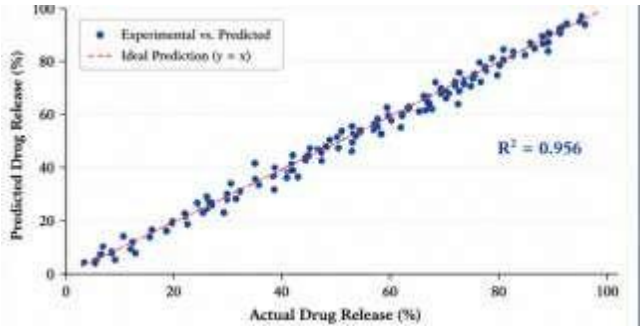

Fig 3. Scatter plot showing agreement between predicted and actual drug release percentages using Gradient Boosting Regression.

The scatter plot demonstrates a strong correlation between predicted and actual release values. Most data points are distributed closely around the ideal prediction line, indicating minimal deviation between experimental observations and model estimates. The absence of significant clustering away from the diagonal line suggests that the model generalizes effectively across different nanocarrier formulations. A small number of prediction deviations were observed at higher release percentages. These discrepancies may be attributed to formulation-specific characteristics that were not fully represented in the available dataset. Nevertheless, the overall distribution confirms the robustness and reliability of the proposed framework.

D. Feature Importance Analysis

Understanding the influence of formulation parameters is important for both predictive modeling and pharmaceutical design. Therefore, feature importance analysis was conducted using the ensemble-based models. The analysis revealed that release duration, encapsulation efficiency, and particle size were the most influential variables affecting prediction outcomes. Release duration exhibited the strongest contribution because cumulative drug release naturally depends on exposure time. Encapsulation efficiency significantly influenced release behavior since it determines the amount of drug retained within the nanocarrier matrix. Particle size also played a major role because smaller particles generally possess larger surface-area-to-volume ratios, which can accelerate drug diffusion. Additional parameters such as polymer concentration, zeta potential, and release medium pH contributed moderately

to prediction performance. Their effects are consistent with pharmaceutical literature, which reports that polymer composition and environmental conditions influence matrix degradation and diffusion pathways.

The identification of influential variables provides practical insights for formulation scientists. By focusing on key parameters during formulation optimization, researchers can improve release characteristics while reducing experimental effort.

E. Discussion and Practical Implications

The obtained results demonstrate that machine learning can serve as a valuable computational tool for intelligent drug delivery research. Traditional formulation development often requires numerous laboratory experiments to evaluate release profiles under varying conditions. Such procedures consume significant time, materials, and financial resources. The proposed framework offers an alternative strategy by enabling rapid prediction of release behavior from formulation characteristics alone. The high accuracy achieved by Gradient Boosting and Random Forest models suggests that machine learning can effectively complement experimental investigations and support early-stage formulation screening. From a practical perspective, predictive models may assist pharmaceutical researchers in identifying promising nanocarrier designs before conducting extensive laboratory testing. This capability can accelerate formulation development, improve resource utilization, and facilitate the design of controlled-release systems. Furthermore, as larger pharmaceutical datasets become available, predictive performance may improve further, enabling broader application in personalized and adaptive drug delivery technologies.

V. CONCLUSION AND FUTURE WORK

This study presented a machine learning-based framework for predicting nanocarrier drug release profiles using physicochemical and formulation parameters. Four regression algorithms, namely Linear Regression, Support Vector Regression, Random Forest, and Gradient Boosting Regression, were evaluated using standardized performance metrics. Experimental results demonstrated that ensemble-based approaches achieved superior predictive accuracy compared with traditional regression methods. Gradient Boosting produced the best overall performance, followed closely by Random Forest. The analysis further revealed that encapsulation efficiency, particle size, and release duration were among the most influential predictors affecting drug release behavior. The proposed framework offers a practical computational approach for supporting nanocarrier formulation development. By enabling accurate prediction of release profiles, machine learning can reduce reliance on extensive laboratory experimentation and accelerate pharmaceutical research workflows. The methodology may also assist researchers in identifying promising formulations during early development stages. Future work will focus on incorporating larger and more diverse nanomedicine datasets, including patient-specific physiological variables. Deep learning architectures and explainable artificial intelligence methods may further enhance predictive performance and interpretability. Integration with digital pharmaceutical platforms could support adaptive and personalized drug delivery systems. The continued convergence of nanotechnology and machine learning is expected to play a significant role in advancing intelligent therapeutic delivery strategies.

REFERENCES

[1] R. Langer, “Drug delivery and targeting,” *Nature*, vol. 392, pp. 5–10, 1998.
[2] S. M. Moghimi, A. C. Hunter, and J. C. Murray, “Nanomedicine: current status and future prospects,” *FASEB Journal*, vol. 19, no. 3, pp. 311–330, 2005.
[3] V. Agrahari and A. K. Mitra, “Nanocarrier fabrication and applications,” *Expert Opinion on Drug Delivery*, vol. 13, no. 9, pp. 1313–1328, 2016.
[4] Y. Barenholz, “Doxil®—The first FDA-approved nano-drug,” *Journal of Controlled Release*, vol. 160, no. 2, pp. 117–134, 2012.
[5] N. A. Peppas and J. J. Sahlin, “A simple equation for drug release,” *International Journal of Pharmaceutics*, vol. 57, pp. 169–172, 1989.
[6] J. S. Mason and M. Goodacre, “Machine learning in drug discovery,” *Drug Discovery Today*, vol. 24, no. 3, pp. 729–740, 2019.

[7] A. Vamathevan et al., "Applications of machine learning in drug discovery and development," *Nature Reviews Drug Discovery*, vol. 18, no. 6, pp. 463–477, 2019.
[8] M. J. Mitchell et al., "Engineering precision nanoparticles for drug delivery," *Nature Reviews Drug Discovery*, vol. 20, pp. 101–124, 2021.
[9] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[10] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.
[11] K. A. Carpenter, H. Huang, and M. K. Machinek, "Machine learning in pharmaceutical formulation development," *AAPS PharmSciTech*, vol. 19, pp. 3058–3068, 2018.
[12] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[13] M. M. Ibrahim and M. A. Al-Harthi, "Artificial intelligence in personalized drug delivery systems," *Drug Delivery and Translational Research*, vol. 12, pp. 1542–1555, 2022.
[14] M. S. Beg, A. R. Ahmad, and S. Jan, "Machine learning approaches in nanomedicine and drug delivery," *Journal of Drug Delivery Science and Technology*, vol. 68, 2022.