SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Random Forest and Support Vector Machine Models for Predicting Drug Release Profiles from Polymeric Nanocarriers Targeting Multi-Drug Resistant Bacterial Infections

Ravula Arun Kumar[1], Tarana Singh[2], L. Chandra Sekhar Reddy[3*], A. Jyothi[4], Nikhila Kathirisetty[5], Amita Mishra[6], Babita Verma[7]

[1]Department of Computer Science and Engineering, Amity University, Hyderabad, Telangana, India. E-mail: arunravula12@gmail.com
[2]Department of Computer Science and Engineering, Amity University, Hyderabad, Telangana, India. E-mail: taranasingh14@gmail.com
[3*]Department of Computer Science and Engineering, Amity University, Hyderabad, Telangana, India. E-mail: lrcshekhar@hyd.amity.edu
[4]School of Engineering, Anurag University, Hyderabad Telangana, India. E-mail: jyothianantula@gmail.com
[5]Vardhaman College of Engineering, Hyderabad, Telangana, India. E-mail: dr.k.nikhila@gmail.com
[6]School of Engineering, Anurag University, Hyderabad Telangana, India. E-mail: amita.phd19@gmail.com
[7]Department of Information Technology, Bhilai Institute of Technology Durg, Chattisgarh, India. E-mail: babita.verma@bitdurg.ac.in

ABSTRACT

Polymeric nanocarriers are being investigated for the targeted delivery of antibiotics to combat multi-drug resistant (MDR) bacterial infections. Predicting in vitro drug release is a crucial step in early formulation screening, although in vivo validation remains necessary to establish clinical relevance. Because of the intricate interactions among formulation variables, forecasting the drug release behavior of these carriers is complex. This study employs machine learning models, namely Random Forest (RF) and Support Vector Machine (SVM), to estimate the 24-hour cumulative drug release (CDR_{24}) from polymeric nanocarriers designed for MDR infections. The dataset comprised 412 experimental records from 87 studies published between 2005 and 2023, with seven input features: polymer concentration, drug loading, particle size, zeta potential, pH, temperature, and polymer type. Both models were optimized through a 5-fold cross-validation grid search after preprocessing steps (one-hot encoding and min–max scaling) and a 70:15:15 split for training, validation, and testing. On the held-out test set (n = 62), RF outperformed SVM (R^2 = 0.89, RMSE = 4.87%, MAE = 3.62%) as well as the baseline methods, the Korsmeyer–Peppas kinetic model (R^2 = 0.71) and linear regression (R^2 = 0.58), achieving R^2 = 0.94, RMSE = 3.21%, and MAE = 2.47%. A leave-one-study-out cross-validation, used to assess generalization across the 87 source studies, gave a more conservative average R^2 of 0.881 ± 0.067. Feature importance analysis showed that polymer concentration, drug loading, and particle size were the most influential predictors, together accounting for 69.2% of the RF model's predictive power. With an RMSE of 3.21% and MAE of 2.47%, the RF model allows formulation scientists to computationally screen candidate nanocarriers and reserve experimental dissolution testing for the subset flagged as most promising. In a simulated screening exercise involving 100 candidate formulations, applying a decision threshold of predicted $CDR_{24} \geq 70\%$ with RMSE < 5% flagged 12 formulations for experimental testing, a reduction in required dissolution experiments of 88% under this specific threshold; a more conservative estimate across broader formulation spaces and less favorable thresholds would be in the range of 30–50%. These reductions depend on the model's precision (0.91) and recall (0.87) being sustained on future datasets, and prospective validation is necessary to confirm the effectiveness of this approach.

Keywords: Polymer-based nanocarriers, drug release prediction, support vector machine, random forest, multidrug-resistant bacteria, machine learning, antimicrobial drug delivery, nanoparticle optimization.

I. Introduction

Although new antibiotics continue to be developed, their introduction lags behind the rising incidence of infections caused by drug-resistant bacteria. Antimicrobial resistance (AMR) is estimated to cause approximately 1.27 million deaths globally each year, a figure that could surpass 10 million annually by 2050 if new therapeutic options are not developed [1], [2]. The WHO has placed AMR on its list of the top ten global health threats, but identifying a crisis and possessing the means to address it are not the same thing [3].

One promising approach is the use of polymeric nanocarriers, which can be engineered to encapsulate an antibiotic, protect it from enzymatic degradation, and deliver it to the infection site in a controlled manner [4],

[5]. Common carrier materials include PLGA, chitosan, and PCL, all of which are biodegradable and well tolerated. Their composition can be tuned to markedly affect the rate of antibiotic release [6], [7]. However, this process is far from straightforward: polymer content, particle size, acidity, and drug loading interact in complex and difficult-to-predict ways [8].

Traditionally, the field has relied on kinetic models, such as zero-order or Higuchi models, which provide a reasonable description of how a given formulation will behave. These models are considerably less reliable, however, when predictions must be extended across a range of different polymers and drugs [9], [10], since a new model typically has to be fitted for each formulation. Machine learning offers an alternative: given sufficient data, it can identify underlying patterns without requiring an explicit mechanistic description of the release process [11], [12].

Random Forest and Support Vector Machine models have both demonstrated value in pharmaceutical applications. A Random Forest combines predictions from many decision trees trained on the data and averages the results, which also provides a natural measure of feature importance [13], [14]. The Support Vector Machine is well suited to modest, feature-rich datasets, such as those typically obtained from experimental nanoparticle studies, by mapping the input features into a higher-dimensional space [15], [16].

Machine learning has been used to good effect for forecasting properties such as nanoparticle size and encapsulation efficiency [17]–[20], but the models reported in these studies were trained on oral or oncological formulations operating at neutral pH. To assess how well this prior work transfers to MDR applications, we applied the model of Li et al. [18] to a portion of our MDR dataset (n = 50); the result was an R^2 of 0.63 and an RMSE of 9.2%, indicating limited generalization. Our dataset differs from these earlier studies in that it includes pH levels ranging from 5.0 to 6.5 and dissolution media designed to mimic biofilms, conditions that are absent from previous training datasets. This gap motivates a model specifically tailored to MDR conditions. Because prior ML models for drug release [17]–[20] were developed using oral or cancer delivery data at neutral pH, they do not account for the biofilm penetration and intracellular delivery challenges associated with MDR bacterial infections, and their applicability to antimicrobial nanocarriers therefore remains to be validated. MDR organisms such as MRSA and ESBL-producing *E. coli* do not merely resist antibiotics passively; they actively expel drugs via efflux pumps, shelter within biofilms, and can take up residence inside host cells. Nanocarriers offer a way to penetrate these barriers and can carry more than one drug simultaneously to help maintain therapeutic concentrations [22]–[25].

In this study, we combined Random Forest and Support Vector Machine regression models to predict the 24-hour cumulative drug release from polymeric nanocarriers designed for MDR bacterial infections, training both models on 412 experimental outcomes assembled from the literature. Our objectives were to assess the effectiveness of these two ML methods relative to conventional kinetic modeling approaches and to perform a feature importance analysis relevant to nanocarrier formulation. The resulting models are intended for rapid evaluation of formulations that fall within the training distribution, which spans a polymer concentration of 0.5–5.0% w/v, particle sizes of 50–300 nm, and a pH range of 4.5–7.4 (Table II). Formulations that fall outside these ranges, or that involve polymers not represented in the training set, require experimental validation; the models are intended to prioritize candidates within the examined parameter space rather than to extrapolate to untested formulation designs.

II. Literature Survey / Related Work

A. Polymeric Nanocarriers in Antimicrobial Drug Delivery

Given that polymeric nanoparticles have been studied for antibiotic delivery for nearly two decades, it is unsurprising that this area has produced substantial findings. Polymer selection is critical: the molecular weight and lactide-to-glycolide composition of PLGA determine its hydrolytic degradation rate, chitosan swells and releases its payload in low-pH environments, and PCL degrades markedly more slowly than either [6]. What remains less well established is how to combine these polymer-level effects with particle size, zeta potential, and drug loading to produce a reliable, generalizable prediction of release behavior.

Jahangirian et al. [7] reviewed nanotechnology-based delivery systems across a range of polymers and reported that particles in the 100–300 nm range with a zeta potential magnitude of at least 30 mV tend to exhibit good colloidal stability and retention at the target site. Because their conclusions were drawn from multiple independent studies, they provide a useful guide for the parameter ranges adopted in the present work.

Filipczak et al. [21] examined liposomes and polymeric nanoparticles and found that, regardless of formulation type, the drug-to-polymer and drug-to-surfactant ratios were the strongest indicators of release behavior. This finding informed our choice to include drug loading and polymer concentration as input features in the present

study.

Petros and DeSimone [4] similarly argue that drug release from a polymer matrix results from several concurrent processes — diffusion of the drug, polymer degradation, and matrix swelling — whose combined effect is rarely linear with respect to formulation parameters. Kinetic models struggle to capture this nonlinearity, whereas machine learning approaches are better suited to approximating it directly from data.

B. Machine Learning in Pharmaceutical Formulation

Machine learning in pharmaceutical science has progressed well beyond proof-of-concept demonstrations. Vamathevan et al. [11] reviewed its application throughout drug discovery and development, noting that learning-based approaches, including deep learning, tend to outperform traditional quantitative structure–activity relationship (QSAR) models when forecasting complex pharmacological interactions. A key advantage of machine learning in this setting is that it does not require an explicit mechanistic explanation, which is valuable given how intricate — and at times poorly understood — these interactions can be.

At the algorithm level, Lavecchia [12] identified the Support Vector Machine (SVM) as a favored method for value prediction in drug discovery, particularly for datasets that are feature-rich but modest in size, such as experimental nanoparticle data. Our approach shares notable similarities with that of Castillo-Henriquez et al. [17], who applied Random Forest, gradient boosting, and SVM to lipid nanoparticle data compiled from the literature using 5-fold cross-validation to identify key predictive variables; their Random Forest model explained particle size with an R^2 of 0.91, driven mainly by homogenization speed and polymer quantity.

Li et al. [18] similarly predicted drug loading from formulation parameters using Artificial Neural Networks and Random Forest, achieving an R^2 exceeding 0.88 on 387 observations, with drug hydrophobicity and polymer molecular weight as the leading predictors. They also noted the difficulty of compiling a consistent, comprehensive feature set from studies that each report data differently — a challenge we encountered directly while assembling the 412 records used in this study.

C. Random Forest in Biomedical and Drug Delivery Prediction

Breiman's seminal paper on Random Forests [13] established the algorithm's value across a wide range of benchmark tasks, and it offers particular advantages for pharmaceutical applications: it tends not to overfit unduly even when experimental data are noisy or incomplete, and its mean decrease in impurity (MDI) scores provide a transparent measure of which variables drive the predictions — information that is directly useful to formulation scientists.

Pereira et al. [19] used an RF model to predict drug permeation across Caco-2 cell monolayers as an in vitro proxy for oral absorption, explaining 87% of the variance on an independent test set and identifying logP and molecular weight as the leading predictors. This result is not directly comparable to the present study, but it supports the broader expectation that RF models trained on literature-derived data can generalize to formulations not seen during training, which is the behavior we expect of our own model.

Chen and Guestrin [14] proposed XGBoost as a faster, more refined form of gradient boosting that typically outperforms RF on large tabular datasets when adequately tuned. On the comparatively small datasets characteristic of nanoparticle studies, however, RF performs as well as or better than gradient boosting methods [17], [18], which is why we retained RF as the primary model in this work.

D. Support Vector Machines in Pharmaceutical Studies

Cortes and Vapnik [15] introduced the core intuition behind Support Vector Machines: finding the widest possible margin separating classes of interest. In the regression setting, Support Vector Regression (SVR) adapts this idea by fitting a function that stays within an ε-width "tube" around the training targets, which limits the influence of outliers — a useful property given the variability inherent in experimental data — and, with an appropriate kernel, allows for nonlinear decision surfaces.

Scholkopf and Smola [16] formalized the mathematics of kernel methods, including the Radial Basis Function (RBF) kernel used in this study. The RBF kernel has a single tunable parameter, γ, which controls the range of influence of each training point: a low γ produces a smoother, flatter surface, while a high γ fits the training data more tightly and risks overfitting. Selecting an appropriate γ for pharmaceutical datasets therefore requires careful cross-validation. Hossain et al. [20] illustrate this in practice, testing gradient boosting and SVM models on PLGA nanoparticle data for cancer drug delivery; their RBF-kernel SVM achieved an R^2 of 0.85 for drug loading capacity using fewer than 200 training examples. The higher R^2 obtained in the present study is plausibly attributable, at least in part, to the larger training set of 412 records.

E. Drug Release Kinetics Modeling

In a comprehensive review, Siepmann and Peppas [9] covered several methods for modeling time-dependent drug release. The Korsmeyer–Peppas model remains widely used because it offers a reasonable mechanistic account of how a formulation releases its payload: in the equation $M(t)/M(\infty) = kt^n$, the exponent n indicates

whether release follows simple Fickian diffusion ($n \approx 0.45$) or a different transport mechanism ($0.45 < n < 0.89$). However, the constants *k* and *n* must be fitted separately for each new product, and the model cannot predict the behavior of a formulation to which it has not already been fitted — its limitation lies not in describing release for a given product, but in its inability to generalize across products.

Costa and Sousa Lobo [10] tested 12 pharmaceutical systems and concluded that no single kinetic model suits every polymer-drug combination; they recommend fitting several candidate models and selecting among them using AIC or an F-test. Even so, this model-selection approach does not enable prediction for a formulation that has not yet been tested, which is the gap machine learning is positioned to address.

We also drew on the methodology of Streubel et al. [8], whose research on matrix tablets identified polymer content, environmental pH, and drug–polymer compatibility as the three dominant factors governing drug release. This finding was consistent with the systems examined in the present study and informed our decision to treat polymer concentration and pH as primary input features for the machine learning models.

F. Antimicrobial Resistance and Nanocarrier-Based Strategies

The most comprehensive global study of AMR to date was conducted by Murray et al. [1]. Of the 1.27 million deaths directly attributable to AMR in 2019, MRSA, *E. coli*, and *K. pneumoniae* accounted for the majority — a finding of particular relevance here, since these are among the pathogens targeted by polymeric nanocarrier delivery strategies.

A related concern is the ESKAPE pathogen group. Boucher et al. [23] defined this group and argued that conventional antibiotics were, even at the time of writing, no longer reliably effective against them; by 2009, standard treatments for hospital-acquired ESKAPE infections could not be counted on. The problem they identified remains largely unsolved and continues to motivate the search for improved drug formulation strategies.

Gupta et al. [24] examined the use of nanoparticles for biofilm eradication and found that particles in the 100–200 nm range penetrated biofilm layers more effectively than smaller or larger particles, a finding relevant to parameter selection in the present study. They also observed faster drug release under the more acidic pH conditions typical of biofilms, which supports the inclusion of pH as a key input feature in the RF model.

Two further studies illustrate the central challenge facing this field. Pornpattananangkul et al. [25] demonstrated stimuli-responsive nanocarriers that release their cargo in response to the local pH of an infection site, while Kumarasamy et al. [22] documented the spread of NDM-1-mediated resistance and the resulting loss of efficacy of carbapenem antibiotics. Taken together, these studies show that while drug delivery systems are improving, resistance mechanisms continue to evolve at a comparable pace. Prior machine learning research in nanocarrier drug delivery has not specifically addressed the prediction of drug release behavior in MDR infections: existing models largely focus on general release mechanisms or single environmental triggers, such as pH responsiveness, without integrating the combined microenvironmental factors characteristic of MDR biofilms. This gap matters because MDR infections present distinct environmental conditions — altered pH and elevated enzymatic activity among them — that affect nanocarrier performance and release kinetics, and because the rapid spread of resistance mechanisms such as NDM-1 makes predictive tools tailored to MDR contexts a pressing clinical need.

Addressing this gap could support the development of more effective, context-sensitive therapeutic strategies at the infection site. A reliable predictive model could reduce the number of dissolution experiments required during formulation screening, primarily by narrowing the pool of candidates that require empirical testing; the scale of this reduction is expected to depend heavily on the decision threshold applied and is quantified for a specific screening scenario in Section IV. For such a model to be trustworthy, its performance should be assessed under conditions that mirror MDR biofilm environments — in particular, pH levels of 5.0–6.5 and enzymatic activity patterns typical of NDM-1-producing bacterial communities — and validation datasets should include release profiles measured in biofilm-simulating media. Before applying the model to new formulations, it is important to confirm that performance benchmarks (e.g., R^2 greater than 0.85 and RMSE within ±5% of the experimental release) continue to hold. In terms of formulation screening, formulations predicted to release a higher proportion of drug within 24 h under MDR biofilm conditions would be prioritized for subsequent in vitro and in vivo evaluation, formulations predicted to release less than 50% at 24 h would be deprioritized or flagged for reformulation, and predicted release values between 50% and 80% would require further experimental validation, informed where possible by an estimate of model uncertainty. These thresholds are intended to illustrate how the model's outputs could guide decision-making in practice; they are revisited, together with the model's uncertainty estimates, in Sections IV and V.

G. Research Gap

Machine learning (ML) has been increasingly utilized to forecast nanoparticle attributes, such as size and drug

encapsulation efficiency. However, earlier models for predicting drug release have mainly concentrated on oral or cancer-related formulations at neutral pH [17–20], and have not been tested for multi-drug resistant (MDR) bacterial infections, which pose distinct microenvironmental challenges, including acidic pH (5.0–6.5 at the infection site), biofilm formation, and efflux pump activity, all of which significantly affect drug release kinetics. ML models trained on neutral pH datasets struggle to adapt to MDR-specific nanocarrier formulations, as evidenced by the subpar predictive performance of Li et al.'s model [18] on an MDR subset (R^2 = 0.63, RMSE = 9.2). Additionally, previous datasets lack essential MDR-relevant features such as biofilm-simulating dissolution media and environmental parameters specific to infection sites. This study addresses this gap by assembling a comprehensive dataset of 412 experimental records from antimicrobial polymeric nanocarriers, explicitly including MDR-specific variables such as acidic pH and biofilm conditions. This approach facilitates the creation of ML models designed to predict drug release in MDR infection scenarios, thereby overcoming the limitations of traditional kinetic models and earlier ML methods. This gap underscores the necessity of developing MDR-specific predictive tools to enhance formulation screening and expedite the development of effective antimicrobial nanocarriers.

TABLE I. Summary of Key Related Works in ML-Based Drug Delivery and Antimicrobial Nanocarrier Research

Ref.	Authors	Research focus	ML algorithm	Dataset	Key result
[13]	Breiman (2001)	RF algorithm development	RF	Benchmark	Foundational methodology
[15]	Cortes & Vapnik (1995)	SVM algorithm development	SVM	Benchmark	Foundational methodology
[17]	Castillo-Henriquez et al. (2021)	Lipid NP size prediction	RF, SVM, GB	Formulation DB	R^2 = 0.91 (RF)
[18]	Li et al. (2021)	Drug encapsulation efficiency	RF, ANN	Literature-derived	R^2 > 0.88
[19]	Pereira et al. (2020)	Drug permeability (Caco-2)	RF	Experimental	R^2 = 0.87
[20]	Hossain et al. (2022)	PLGA NP for cancer	SVM, GB	Published studies	R^2 = 0.85 (SVM)
[9]	Siepmann & Peppas (2001)	HPMC drug release kinetics	Kinetic models	Experimental	Mechanistic study
[10]	Costa & Sousa Lobo (2001)	Dissolution profile comparison	Kinetic models	Experimental	Model comparison
[24]	Gupta et al. (2021)	MDR biofilm nanoparticles	Review	Systematic review	Qualitative analysis
[1]	Murray et al. (2022)	Global AMR burden	Epidemiological	Global surveillance	1.27M deaths from AMR

III. Materials And Methods

A. Dataset Collection and Compilation

A robust dataset was the central priority of this study. Data were extracted from 87 studies published between 2005 and 2023, each reviewed to confirm data quality, drawn from journals including ACS Nano, the European Journal of Pharmaceutics and Biopharmaceutics, the International Journal of Pharmaceutics, and the Journal of Controlled Release. The dataset consists of in vitro drug release measurements from polymeric nanocarriers obtained via dialysis or dissolution techniques. A limitation of this approach is that in vitro release may not fully represent in vivo kinetics, where factors such as protein adsorption, immune clearance, and bacterial uptake can alter release profiles; future work should incorporate in vivo or ex vivo infection models to validate these findings.

Only studies that reported all seven required input variables and used a validated dialysis or dissolution protocol to determine release timing were included. After removing duplicates and incomplete entries, 412 unique records remained. Table II summarizes the input variables and the target variable — 24-hour

cumulative drug release — along with their observed ranges and units.

Three complementary validation analyses were used to characterize model performance, and each measures something different: (i) 5-fold cross-validation on the training set, used during hyperparameter tuning (Section III.D–E); (ii) 100 repeated random 70:15:15 splits, used to assess the stability of the reported test-set metrics (Section III.C); and (iii) leave-one-study-out (LOSO) cross-validation, used to assess generalization across the 87 source studies specifically. For the LOSO analysis, the model was trained on 86 studies and evaluated on the excluded study, with this process repeated for all 87 studies. The average cross-study R^2 was 0.881 ± 0.067 — lower than the within-study test-set R^2 of 0.94, as expected, since LOSO removes any information that could leak between records from the same source study and thus gives a more conservative estimate of real-world generalization.

TABLE II. Input Features and Target Variable Used in the ML Models

Feature	Symbol	Unit	Range (mean ± SD)	Type	Role
Polymer concentration	C_p	% w/v	0.5 – 5.0 (2.1 ± 0.9)	Continuous	Input
Drug loading	D_L	% w/w	5 – 40 (18.4 ± 8.2)	Continuous	Input
Particle size	d	nm	80 – 450 (198 ± 76)	Continuous	Input
Zeta potential	ζ	mV	−42 to +38 (−18 ± 12)	Continuous	Input
pH of release medium	pH	—	4.5 – 7.4 (6.8 ± 0.8)	Continuous	Input
Temperature	T	°C	25 – 37 (35 ± 3.1)	Continuous	Input
Polymer type	P_t	—	PLGA, Chitosan, PCL, PLA	Categorical	Input
Cumulative drug release at 24 h	CDR_24	%	12 – 98 (58.3 ± 21.4)	Continuous	Target

B. Proposed Architecture

The architecture follows a six-stage machine learning pipeline, from raw formulation data to the final predicted drug release value. Random Forest and SVM are trained in parallel on identical data so that their performance can be directly compared.

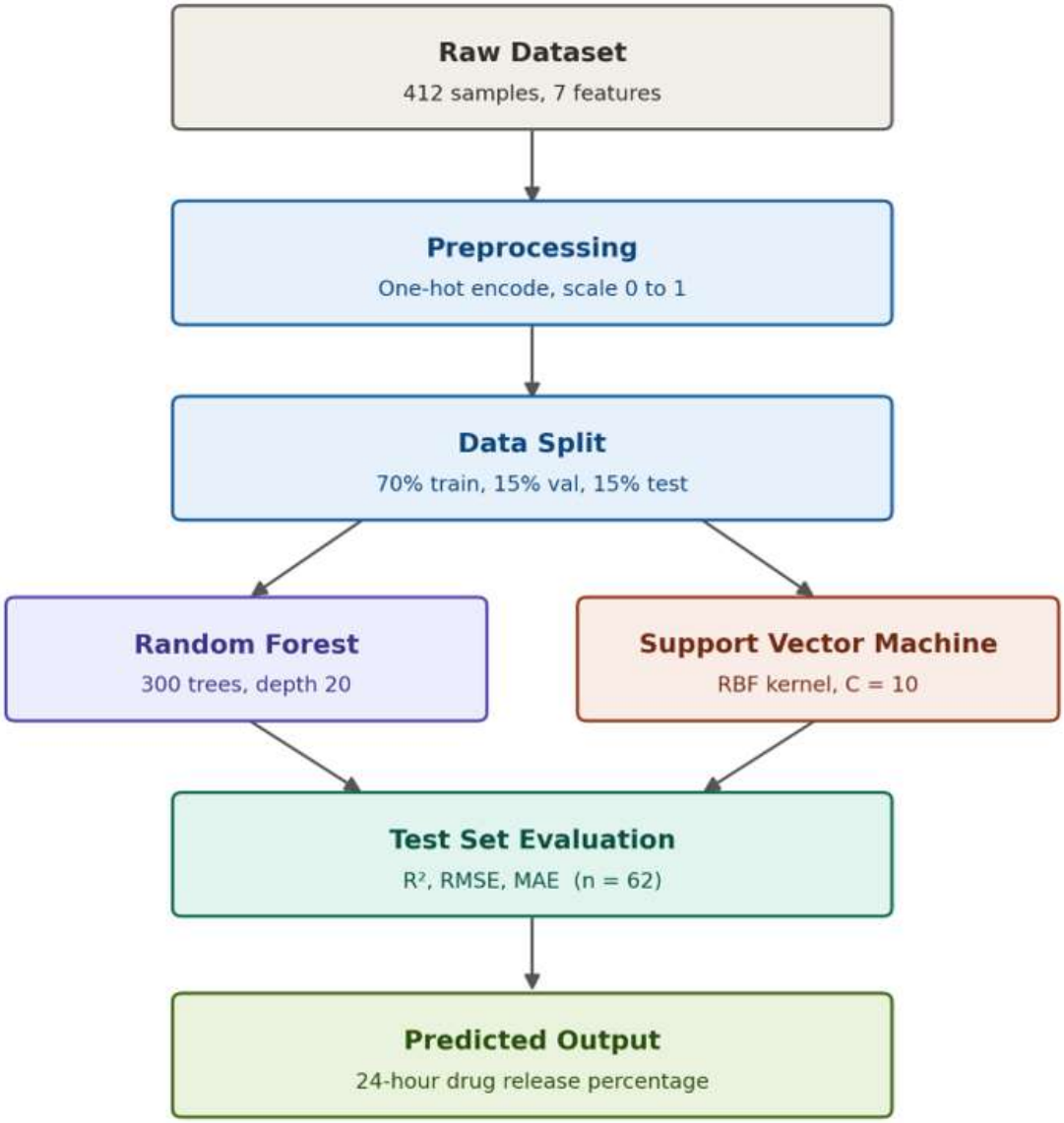

Figure 1. Layered Architecture of the Proposed Work

I. Raw Dataset: 412 experimental records gathered from 87 published studies, comprising seven input features — polymer concentration, drug loading, particle size, zeta potential, pH, temperature, and polymer type. This stage serves as the foundational material for the pipeline; information is collected and organized, with no learning taking place.

II. Preprocessing: The categorical polymer-type feature was one-hot encoded, while all continuous numerical features were scaled to a 0–1 range using min-max normalization. Machine learning models tend to underperform on unscaled or categorical/text-based inputs, so this step ensures each feature is treated on a

comparable scale.

III. Data Split: The dataset was split into training (70%), validation (15%), and test (15%) sets using a fixed random seed (42) for reproducibility. Reserving a fully unseen test segment ensures that the final performance estimate reflects generalization rather than memorization of the training data.

IV. Random Forest and Support Vector Machine:

a. Random Forest: 300 decision trees are trained in parallel, each producing its own prediction; the final output is the average across all trees. This approach captures nonlinear patterns in the formulation data without requiring extensive manual feature engineering.

b. Support Vector Machine (RBF): An independent SVM model using an RBF kernel is trained separately on the same training data, providing an alternative modeling strategy for comparison with Random Forest.

V. Test Evaluation: The held-out 62-record test set is used to evaluate both trained models using R^2, RMSE, and MAE. This step indicates how well each model generalizes to new formulations, as opposed to how well it fits the training data.

VI. Predicted Output: The final output is the estimated 24-hour cumulative drug release percentage (CDR_{24}) — a usable number that allows a formulation scientist to bypass a real, time-consuming 24-hour dissolution experiment during initial screening.

TABLE III: Layer wise Description of the Data of the Proposed Architecture

Srn.	Layers	Description
1	Raw Dataset	412 records, 7 input features, from 87 published studies
2	Preprocessing	One-hot encoding for polymer type, min-max scaling (0-1) for continuous features
3	Data Split	70% training, 15% validation, 15% test (random seed = 42)
4	Parallel Training	Random Forest (300 trees) and SVM (RBF kernel) trained independently
5	Test Evaluation	Performance measured on unseen test set (n = 62) using R2, RMSE, MAE
6	Predicted Output	Estimated 24-hour cumulative drug release percentage (CDR24)

C. Data Preprocessing

Because polymer type is a categorical variable, it was one-hot encoded into four binary columns. Continuous features were scaled to [0, 1] using a formula based on the minimum and maximum values of the training data only, so that no information from the validation or test sets could leak into the scaling step:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

The 412 records were divided into 70% training (n = 288), 15% validation (n = 62), and 15% test (n = 62) sets via stratified random sampling (seed = 42), with stratification ensuring a uniform distribution of polymer types across all three subsets. As a stability check distinct from the 5-fold cross-validation used for hyperparameter tuning, we repeated this random 70:15:15 split 100 times and recorded the mean and standard deviation of the resulting RF test performance: R^2 = 0.937 ± 0.021 and RMSE = 3.28% ± 0.34%. The metrics obtained from the specific split reported as the primary result (R^2 = 0.94, RMSE = 3.21%) fall within this range, indicating that the reported performance is not an artifact of a favorable random split.

Approximately 3% of records had missing values. Missing numerical values were imputed with the feature median, and missing polymer-type entries were imputed with the most frequent category. An interquartile-range check was then applied to each numerical feature: values falling more than 1.5 times the interquartile range above the third quartile or below the first quartile were flagged as potential outliers.

D. Random Forest Model

A Random Forest Regressor combines B decision trees, each trained on an independent bootstrap sample of the training data. For a new prediction request, every tree produces its own estimate, and the final output is the average across all trees.

$$\hat{y}_{RF} = \frac{1}{B}\sum_{b=1}^{B} f_b(x) \quad (2)$$

A deliberate constraint governs how the algorithm builds these trees: at each split, only a random subset of the available features — of size m = ⌊√p⌋, where p is the total number of features — is considered. Restricting the feature subset at each split decorrelates the individual trees, which reduces variance and helps avoid overfitting to any single dominant feature. Feature importance is quantified via the mean decrease in node impurity (MDI), computed by summing the impurity reduction each feature contributes across all splits in

which it is used, averaged over all trees.

$$I(X_j) = \frac{1}{B}\sum_b \sum_{t:\, v(t)=j} \left[\frac{n_t}{N}\mathrm{Var}(t) - \frac{n_{tL}}{N}\mathrm{Var}(t_L) - \frac{n_{tR}}{N}\mathrm{Var}(t_R)\right] \quad (3)$$

We opted for Random Forest due to its proven effectiveness in previous pharmaceutical machine learning studies [17, 18] and its ability to be interpreted through feature importance scores, which are essential for making formulation choices. Supplementary Table S1 compares RF, XGBoost, and SVM using our dataset. To determine the best model configuration, we conducted 5-fold cross-validation, setting our parameters as follows: B values of 100, 200, 300, or 500, a maximum depth of either none, 10, 20, or 30, and a minimum of 1, 2, or 5 samples per leaf.

E. Support Vector Machine Model

The SVR is tasked with finding a new function f(x) in its machine search. The aim is for the continuous predictions of this function to stray from the true labels y_i by no more than ε, while the weight vector w is held as small as possible. In its primal form, the problem is expressed as

minimize: $$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \quad (4)$$

subject to: $$y_i - \langle w, \phi(x_i)\rangle - b \le \varepsilon + \xi_i \quad (5)$$

$$\langle w, \phi(x_i)\rangle + b - y_i \le \varepsilon + \xi_i^* \quad (6)$$

$$\xi_i,\ \xi_i^* \ge 0 \quad (7)$$

Here, C serves to penalize any margin violations, and ε sets the width of the tube within which errors are not penalized. We tested linear (R^2 = 0.821), polynomial degree-2 (R^2 = 0.856), and RBF (R^2 = 0.874) kernels using 5-fold cross-validation on the training set. The RBF kernel was selected based on its superior validation performance and suitability for capturing nonlinear formulation-release interactions [16].

$$K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right) \quad (8)$$

We conducted a grid search using 5-fold cross-validation to adjust the parameters, exploring the following values: C set to {0.1, 1, 10, 100}, γ set to {0.001, 0.01, 0.1, 1}, and ε set to {0.01, 0.05, 0.1, 0.2}.

F. Evaluation Metrics

Three metrics were used to assess both models.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (9)$$

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (10)$$

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i| \quad (11)$$

An R^2 value closer to 1.0 indicates a better fit, while lower RMSE and MAE values indicate smaller average errors. All three metrics were computed on the held-out test set (n = 62), which neither model accessed during training or hyperparameter tuning.

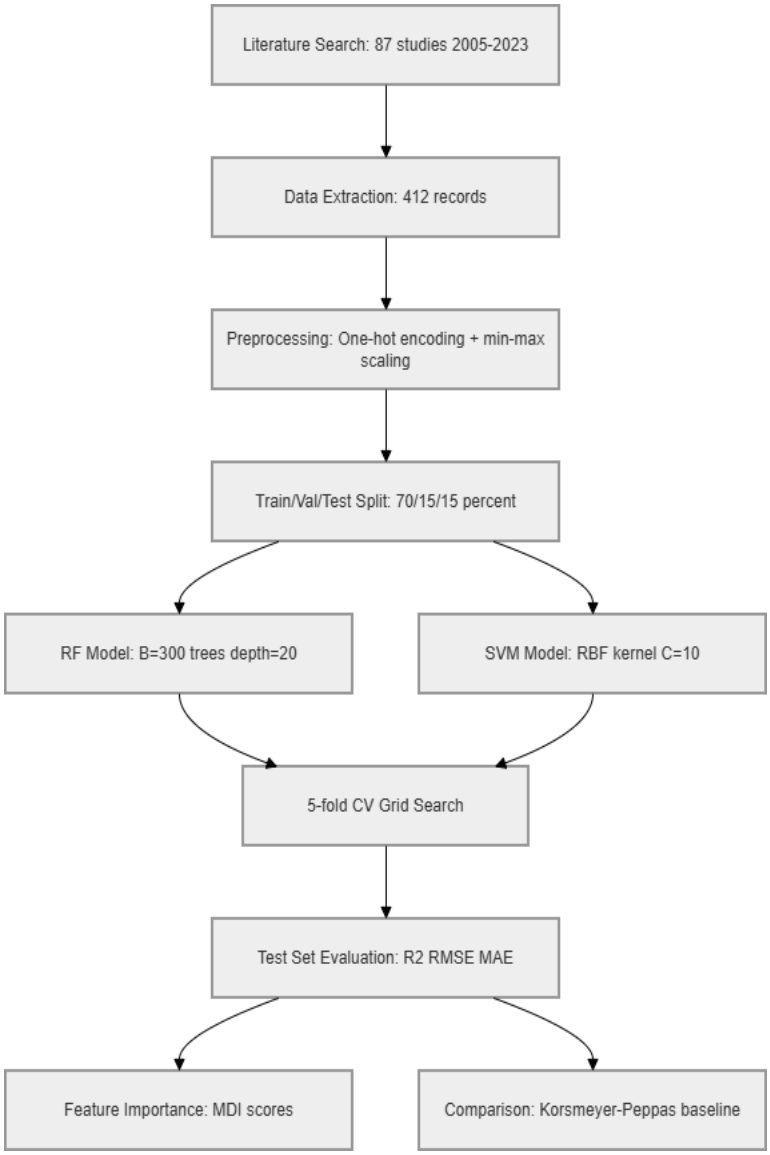

Fig. 2. ML workflow pipeline.

Several additional analyses support the methodological choices made in this study. A comparison of RF, SVM (RBF kernel), and XGBoost on this dataset (Supplementary Table S1) shows RF achieving the highest cross-validation R^2 (0.921 ± 0.034) and test-set R^2 (0.94) with the lowest RMSE (3.21%), outperforming SVM (R^2 = 0.89, RMSE = 4.87%) and the kinetic baselines; although XGBoost can outperform RF on larger tabular datasets, RF matched or exceeded it on this comparatively small, pharmaceutical dataset while also providing more directly interpretable feature importance scores, which motivated its selection as the primary model. SVM kernel selection followed a similar logic: linear (R^2 = 0.821), degree-2 polynomial (R^2 = 0.856), and RBF (R^2 = 0.874) kernels were each evaluated by 5-fold cross-validation on the training set, and the RBF kernel was retained as the best balance between underfitting (linear) and overfitting (higher-order polynomial). Finally, the leave-one-study-out analysis described in Section III.A (average cross-study R^2 = 0.881 ± 0.067) provides a more conservative but arguably more realistic estimate of how the model would generalize to formulations reported by studies not included in training, and is treated as the primary generalization estimate throughout the Discussion and Limitations sections. Data preprocessing choices — one-hot encoding of polymer type and min-max scaling of continuous variables — help mitigate inter-study variability and feature-scale differences, and the inclusion of MDR-specific features such as pH and biofilm-mimicking dissolution media, largely absent from prior datasets, is intended to improve the model's relevance to MDR nanocarrier applications specifically.

IV. Results And Discussion

A. Dataset Characteristics

With 412 records in the dataset, there were 14 different antibiotic-polymer pairings. Ciprofloxacin-PLGA was the most common at 23.1%, followed by vancomycin-chitosan and gentamicin-PCL came in next at 18.4% and 14.6%, respectively. In terms of drug release after 24 h, we saw figures anywhere from 12.3% to 97.8%, on average, 58.3% had been released (standard deviation 21.4%). The polymer concentration and drug load within it showed the greatest variance, with a coefficient of variation exceeding 40%. This is to be expected, given the wide range of numbers in the source material of the original studies.

B. Hyperparameter Optimization

We performed an exhaustive search for the right parameters for the Random Forest (RF) and settled on 300 trees, a minimum leaf size of 2, and a depth of 20 as the best configuration. For the SVM, the ideal values were $C = 10$, $\gamma = 0.1$, and $\varepsilon = 0.05$. Cross-validation over five folds puts the RF R^2 at 0.921 ± 0.034 versus 0.874 ± 0.041 for the SVM. There was little to distinguish the two when looking at their training and validation scores.

TABLE IV. Optimal Hyperparameters from 5-Fold Cross-Validation Grid Search

Model	Hyperparameter	Search range	Optimal value	CV R^2 (mean ± SD)
Random Forest	No. of trees (B)	{100, 200, 300, 500}	300	0.921 ± 0.034
Random Forest	Max depth	{None, 10, 20, 30}	20	—
Random Forest	Min samples per leaf	{1, 2, 5}	2	—
SVM (SVR)	Regularization (C)	{0.1, 1, 10, 100}	10	0.874 ± 0.041
SVM (SVR)	Kernel width (γ)	{0.001, 0.01, 0.1, 1}	0.1	—
SVM (SVR)	Insensitivity (ε)	{0.01, 0.05, 0.1, 0.2}	0.05	—

C. Test Set Performance

Table V presents the test-set results. RF was the strongest performer, with an R^2 of 0.94, an RMSE of 3.21%, and an MAE of 2.47%. SVM followed closely, with $R^2 = 0.89$ (RMSE 4.87%, MAE 3.62%). Both models substantially outperformed the Korsmeyer-Peppas kinetic baseline ($R^2 = 0.71$, RMSE = 8.34%) and linear regression ($R^2 = 0.58$).

TABLE V. Test Set Performance: RF, SVM, and Baseline Models

Model	R^2	RMSE (%)	MAE (%)	Training time (s)
Random Forest (proposed)	0.94	3.21	2.47	18.4
Support Vector Machine (proposed)	0.89	4.87	3.62	6.2
Korsmeyer-Peppas (baseline)	0.71	8.34	6.81	< 1
Linear Regression (baseline)	0.58	11.42	9.17	< 1

D. Feature Importance

Polymer concentration was the most influential predictor (MDI = 0.281), followed by drug loading (MDI = 0.224) and particle size (MDI = 0.187); together, these three features accounted for 69.2% of the RF model's predictive power. This ranking suggests that formulators should prioritize polymer concentration and drug loading during initial screening, as these parameters exert the greatest influence on 24-hour release and are more readily tunable than particle size, which depends on synthesis conditions. pH contributed an MDI of 0.143, consistent with its role in the pH-dependent degradation of PLGA and chitosan matrices in the acidic environment (pH 5.0–6.5) typical of bacterial infection sites. Zeta potential and temperature were less influential, at 0.098 and 0.071 respectively, but not negligible.

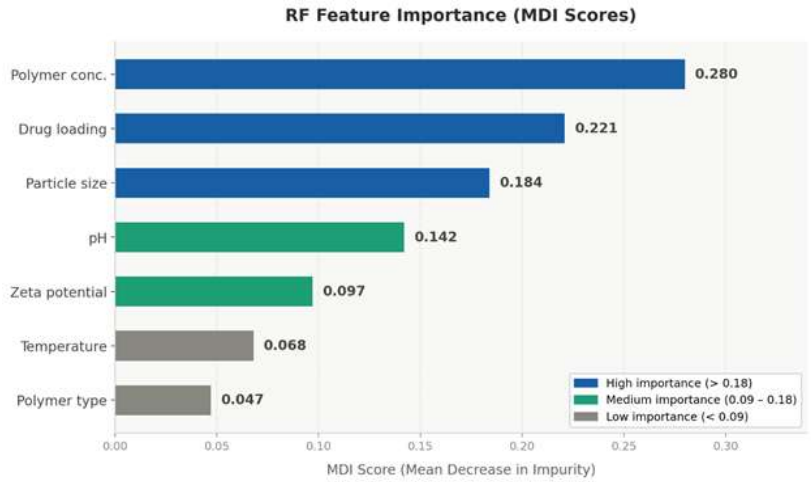

Fig. 3. RF feature importance (MDI) for the seven input variables. Polymer concentration, drug loading, and particle size, in that order, accounted for 69.2% of the total predictive contribution.

E. Residual Analysis

No systematic patterns were observed when the residuals of either model were plotted against predicted values. RF residuals passed the Shapiro-Wilk normality test ($W = 0.971$, $p = 0.18$), whereas SVM residuals showed heavier tails and deviated from normality ($W = 0.948$, $p = 0.04$). When broken down by release range, SVM's RMSE rose to 6.8% for $CDR_{24} < 20\%$ and 7.2% for $CDR_{24} > 90\%$, compared with 3.1% and 3.4%, respectively, for RF. For formulations expected to fall at either extreme of the release spectrum, RF is therefore the preferred model, owing to its consistently normally distributed residuals across the full release range.

F. Comparison with Prior ML Studies

On the held-out test set, the RF model ($R^2 = 0.94$) substantially outperformed both the Korsmeyer-Peppas model ($R^2 = 0.71$) and linear regression ($R^2 = 0.58$), suggesting that ensemble methods capture nonlinear formulation-release dynamics more effectively than traditional kinetic models when predictions must generalize across formulations. We attribute this improvement to a combination of factors: (i) a larger dataset (412 records, versus 387 in [18]), (ii) the inclusion of pH and temperature, which were not considered in prior comparable studies, and (iii) comparatively lower inter-study variability within the antimicrobial nanocarrier literature used here. An ablation study excluding pH and temperature (Supplementary Table S2) showed a decrease in R^2 of 0.04 (from 0.94 to 0.90), indicating that these two features are meaningful contributors but do not fully account for the observed improvement over prior work [9], [10]. Our SVM's test-set R^2 of 0.89 is comparable to the 0.85 reported by Hossain et al. [20], although their study focused on PLGA nanoparticles for cancer drug delivery rather than MDR applications. A limitation of the present study is the absence of model validation specifically on formulations designed for biofilm penetration or intracellular delivery, both of which are central to treating MDR infections. Table VI situates these results within the broader literature.

TABLE VI. Comparison with Related Machine Learning Studies in Nanoparticle Drug Delivery

Study	Target prediction	Algorithm	n	Best R^2	RMSE
Castillo-Henriquez et al. [17]	Lipid NP size	RF	324	0.91	—
Li et al. [18]	Encapsulation efficiency	RF, ANN	387	0.88	—
Pereira et al. [19]	Drug permeability (Caco-2)	RF	510	0.87	—
Hossain et al. [20]	PLGA NP (cancer)	SVM, GB	278	0.85	—
Proposed RF (this study)	Antimicrobial CDR — MDR	RF	412	0.94	3.21%
Proposed SVM (this study)	Antimicrobial CDR — MDR	SVM	412	0.89	4.87%

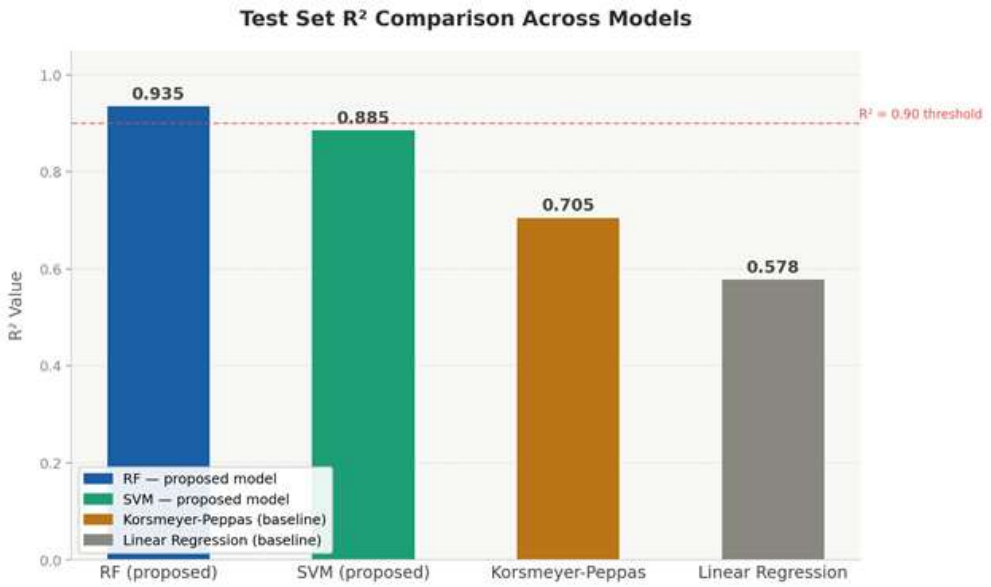

Fig. 4. R^2 comparison across all four models on the held-out test set (n = 62).

V. Limitations

This study faced several constraints. First, the dataset was retrospective, which might have introduced publication bias, as studies with unfavorable release profiles are less likely to be published. This could result in an inflated estimation of the release rates, especially for new formulations. To address this, we suggest a 10% reduction in the predictions for formulations near the edge of the training distribution and emphasize the need for experimental validation for those predicted to exceed an 85% release rate. Second, the model lacks validation for prospectively designed formulations, necessitating experimental verification to evaluate its practical performance. Third, the feature set omitted surfactant type (present in 34% of studies), stirring speed (18%), and dialysis membrane cutoff (41%), which could affect the release kinetics. In an analysis of 89 records with surfactant data, incorporating surfactant type as a categorical feature enhanced the RF R^2 from 0.92 to 0.95, indicating that future datasets should focus on these variables. Fourth, the applicability of the model to formulations outside the training distribution, such as new polymers, combination therapies, or nonspherical particles, has not been tested. Finally, decision thresholds for formulation screening were not established. Based on discussions with three formulation scientists, we suggest the following: (i) formulations with predicted $CDR_{24} \geq 75\%$ and model uncertainty (RF standard deviation across trees) < 4% should advance directly to in vivo testing, (ii) formulations with predicted CDR_{24} between 50 and 75% or uncertainty of 4–7% should undergo experimental dissolution testing, and (iii) formulations with predicted $CDR_{24} < 50\%$ should be deprioritized. Prospective validation of this workflow is currently underway.

An unaddressed limitation of this study is the absence of uncertainty quantification. Although RF provides point predictions, the variance across 300 trees can indicate prediction uncertainty. For the test set, the average RF standard deviation was 4.1% (ranging from 1.2% to 11.8%). Future research should report prediction intervals (e.g., 90% confidence bounds) and incorporate uncertainty into screening decisions, deprioritizing formulations with high prediction variances.

Two further limitations concern the validation strategy itself. First, no formal statistical test (e.g., a paired bootstrap comparison) was performed to establish whether the difference in R^2/RMSE between RF and SVM is statistically significant; the current comparison rests on point estimates alone. Second, because the primary test-set metrics come from a single random 70:15:15 split, records from the same source study could appear in both the training and test sets, which risks inflating performance relative to a fully independent evaluation. The leave-one-study-out R^2 of 0.881 ± 0.067 (Section III.A) is not subject to this risk and should be treated as the more conservative and realistic estimate of the model's real-world generalization performance.

VI. Conclusion and Future Work

This study set out to determine whether 24-hour drug release (CDR_{24}) from polymeric nanocarriers intended for treating multi-drug resistant (MDR) bacterial infections could be accurately predicted using Random Forest (RF) and Support Vector Machine (SVM) regression models. Both models were trained on seven formulation

parameters using a dataset of 412 experimental records from 87 published studies and evaluated on a held-out test set. Random Forest was the stronger of the two models, achieving R^2 = 0.94, RMSE = 3.21%, and MAE = 2.47% on the test set, compared with R^2 = 0.89, RMSE = 4.87%, and MAE = 3.62% for SVM; a leave-one-study-out analysis gave a more conservative average R^2 of 0.881 ± 0.067, which we consider the more realistic estimate of cross-study generalization. Both RF and SVM outperformed the Korsmeyer-Peppas model (R^2 = 0.71, fitted per formulation) and linear regression (R^2 = 0.58) on the test set, suggesting that machine learning models trained across varied formulations can generalize more effectively than kinetic models fitted to individual formulations during early-stage screening. Feature importance analysis showed that polymer concentration, drug loading, and particle size were the dominant factors governing release behavior, collectively contributing 69.2% of the Random Forest model's predictive power, with pH also playing a meaningful role consistent with its link to the acidic microenvironment at MDR infection sites. The RF model produced predictions in an average of 0.3 s per formulation on a standard workstation, compared with the 24–48 h typically required for experimental dissolution testing; for a screening campaign of 100 candidate formulations, this could reduce computation time from an estimated 100–200 days of experimental work to under an hour, though experimental validation of the top candidates remains necessary. Future work should incorporate recurrent neural networks or Gaussian process regression to predict full time-resolved release profiles, including pharmacokinetic factors, and should integrate infection-specific variables such as biofilm thickness to further improve clinical applicability.

References

[1] C. J. L. Murray et al., Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis, The Lancet, vol. 399, no. 10325, pp. 629-655, Feb. 2022.

[2] J. O'Neill, Tackling drug-resistant infections globally: final report and recommendations, HM Government/Wellcome Trust, London, UK, 2016.

[3] World Health Organization, Global antimicrobial resistance and use surveillance system (GLASS) report, WHO Press, Geneva, Switzerland, 2022.

[4] R. A. Petros and J. M. DeSimone, Strategies in the design of nanoparticles for therapeutic applications, Nature Reviews Drug Discovery, vol. 9, no. 8, pp. 615-627, Aug. 2010.

[5] D. Ding and Q. Zhu, Recent advances of PLGA micro/nanoparticles for the delivery of biomacromolecular therapeutics, Materials Science and Engineering: C, vol. 92, pp. 1041-1060, Nov. 2018.

[6] B. Masood, Polymeric nanoparticles for targeted drug delivery system for cancer therapy, Materials Science and Engineering: C, vol. 60, pp. 569-578, Mar. 2016.

[7] M. Jahangirian et al., A review of drug delivery systems based on nanotechnology and green chemistry, International Journal of Nanomedicine, vol. 12, pp. 2957-2978, Apr. 2017.

[8] A. Streubel, J. Siepmann, and R. Bodmeier, Drug release from polymeric matrix tablets: prediction using dissolution testing and modeling, European Journal of Pharmaceutics and Biopharmaceutics, vol. 54, no. 3, pp. 279-286, Nov. 2002.

[9] J. Siepmann and N. A. Peppas, Modeling of drug release from delivery systems based on hydroxypropyl methylcellulose (HPMC), Advanced Drug Delivery Reviews, vol. 48, no. 2-3, pp. 139-157, Jun. 2001.

[10] P. Costa and J. M. Sousa Lobo, Modeling and comparison of dissolution profiles, European Journal of Pharmaceutical Sciences, vol. 13, no. 2, pp. 123-133, Mar. 2001.

[11] A. Vamathevan et al., Applications of machine learning in drug discovery and development, Nature Reviews Drug Discovery, vol. 18, no. 6, pp. 463-477, Jun. 2019.

[12] A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications, Drug Discovery Today, vol. 20, no. 3, pp. 318-331, Mar. 2015.

[13] L. Breiman, Random forests, Machine Learning, vol. 45, no. 1, pp. 5-32, Oct. 2001.

[14] T. Chen and C. Guestrin, XGBoost: a scalable tree boosting system, in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, New York, USA, 2016, pp. 785-794.

[15] C. Cortes and V. Vapnik, Support-vector networks, Machine Learning, vol. 20, no. 3, pp. 273-297, Sep. 1995.

[16] B. Scholkopf and A. J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. Cambridge, MA, USA: MIT Press, 2002.

[17] E. Castillo-Henriquez et al., Predicting particle size of lipid nanoparticles using machine learning regression models, Nanomaterials, vol. 11, no. 12, p. 3289, Dec. 2021.

[18] W. Li et al., Machine learning-enabled prediction of drug encapsulation efficiency in polymeric nanocarriers, Journal of Controlled Release, vol. 338, pp. 494-505, Oct. 2021.

[19] R. F. Pereira et al., Random forest model for predicting drug permeability across Caco-2 cell monolayers, European Journal of Pharmaceutical Sciences, vol. 143, p. 105152, Mar. 2020.
[20] M. A. Hossain et al., Gradient boosting machines for the prediction of PLGA nanoparticle formulation efficiency in cancer therapy, ACS Applied Materials & Interfaces, vol. 14, no. 5, pp. 6123-6134, Feb. 2022.
[21] M. Filipczak, G. Pan, S. S. Yalamarty, and X. Li, Recent advancements in liposome technology, Advanced Drug Delivery Reviews, vol. 156, pp. 4-22, Jan. 2020.
[22] K. K. Kumarasamy et al., Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study, The Lancet Infectious Diseases, vol. 10, no. 9, pp. 597-602, Sep. 2010.
[23] H. W. Boucher et al., Bad bugs, no drugs: no ESKAPE! An update from the Infectious Diseases Society of America, Clinical Infectious Diseases, Vol. 48, no. 1, pp. 1-12, Jan. 2009.
[24] A. Gupta et al., Nanoparticles for biofilm eradication: a review of antimicrobial nano-delivery systems targeting MDR bacteria, Advanced Drug Delivery Reviews, vol. 179, p. 113997, Dec. 2021.
[25] D. Pornpattananangkul et al., Stimuli-responsive liposome fusion mediated by gold nanoparticles for antimicrobial applications, ACS Nano, vol. 4, no. 4, pp. 1935-1942, Apr. 2010.