Machine Learning based approach for the design and release kinetics prediction of the PLGA-based Drug Delivery Systems

APURBA DEBNATH

20BT8019

MAY 2024

Machine Learning based approach for the design and release kinetics prediction of the PLGA–based Drug Delivery Systems

Thesis is submitted in partial fulfillment.

of the requirements for the degree of

BACHELOR OF TECHNOLOGY

in

BIOTECHNOLOGY

By

APURBA DEBNATH

20BT8019

Under the guidance of

Prof. Dalia Dasgupta Mandal

Professor

Department of Biotechnology

National Institute of Technology

Durgapur, India

MAY 2024

DEPARTMENT OF BIOTECHNOLOGY
NATIONAL INSTITUTE OF TECHNOLOGY
DURGAPUR, INDIA

DECLARATION

I the undersigned declare that the thesis work entitled "**Machine Learning based approach for the design and release kinetics prediction of the PLGA-based Drug Delivery Systems** ", submitted towards partial fulfillment of requirements for the award of the degree in **Bachelor of Technology** in **Biotechnology** is my original work and this declaration does not form the basis for award of any degree or any similar title to the best of my / our knowledge.

Durgapur
May 2024

Name of the Student
Roll No.

DEPARTMENT OF BIOTECHNOLOGY
NATIONAL INSTITUTE OF TECHNOLOGY
DURGAPUR, INDIA

CERTIFICATE OF RECOMMENDATION

This is to certify that the thesis entitled "**Machine Learning based approach for the design and release kinetics prediction of the PLGA-based Drug Delivery Systems** ", submitted by **Apurba Debnath** of Department of Biotechnology, National Institute of Technology, Durgapur, in partial fulfillment of the requirements for the award of the degree in **Bachelor of Technology** in **Biotechnology** is a bonafide record of work carried out by him under my guidance during the academic year 2023 – 2024.

Professor and Head
Department of Biotechnology
National Institute of Technology
Durgapur

Prof. Dalia Dasgupta Mandal
Professor
Department of Biotechnology
National Institute of Technology
Durgapur

DEPARTMENT OF BIOTECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY

DURGAPUR, INDIA

CERTIFICATE OF APPROVAL

This is to certify that we have examined the thesis entitled "**Machine Learning based approach for the design and release kinetics prediction of the PLGA-based Drug Delivery Systems"**, submitted by **Apurba Debnath** and hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance in partial fulfillment of the requirements for the award of the degree in **Bachelor of Technology** in **Biotechnology** for which it has been submitted. It is to be understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed, or conclusion drawn therein but approve the thesis only for the purpose for which it is submitted.

Examiners:

Name	**Signature**

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor ***Prof. Dalia Dasgupta Mandal,*** *Professor, Department of Biotechnology, National Institute of Technology Durgapur* for enlightening me the first glance of research, and for her patience, motivation, enthusiasm, and immense knowledge. Her inspiring guidance, systematic approach, sensible criticisms, and close support throughout the course of this project work helped me in overcoming the problems at many critical stages of the assignment leading and enabling towards a successful accomplishment of this project.

I am also very grateful to all the professors of the Biotechnology department for their continuous assistance and encouragement which served as pillars of strength throughout this journey. And dep thankful to all my classmates and friends for their love and support.

Finally, I feel great reverence for all my family members and the Almighty, for their blessings and for being a constant source of encouragement.

Durgapur

May 2024

APURBA DEBNATH

Roll No.- 20BT8019

Department of Biotechnology

NIT Durgapur

ABSTRACT

Biodegradable polymer-based drug delivery systems (DDS), particularly those using Poly (lactic-co-glycolic acid) (PLGA), have shown significant promise in achieving sustained and controlled drug release to targeted sites within the body, maximizing therapeutic response while minimizing adverse side effects. However, designing such systems and predicting release kinetics are highly complex and resource-intensive tasks, as they depend on the intricate interplay of various physicochemical properties of polymers, drugs, and their interactions within the matrix. Traditional trial-and-error methods are both time-intensive and expensive, involving manual exploration of parameter spaces and the rational design of release profiles. This process often takes years to fine-tune parameters for achieving optimal therapeutic efficacy. To address these challenges, the study introduces a computational approach leveraging machine learning (ML) models trained on historical drug release profiles to predict the release rate in advance and identify key factors influencing drug release. The trained models, particularly tree-based ML models like Random Forest and XGBoost, exhibited strong predictive performance in modeling release kinetics and successfully identified key physicochemical factors. Altogether, the findings offer actionable insights into the design of PLGA-based DDS and identify critical physicochemical parameters influencing drug release, paving the way for more efficient and cost-effective development of advanced DDS.

Contents

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1. INTRODUCTION

Poly (lactic-co-glycolic acid) (PLGA), one of the most widely used biodegradable polymers in the field of drug delivery systems in modern medicine. Its unique properties arise from the combination of lactic acid and glycolic acid, which allows for a broad range of variations in its chemical structure. By adjusting the ratio of these two monomers and the molecular weight of the polymer, PLGA can be tailored to achieve specific degradation rates and drug release profiles, making it highly versatile for different therapeutic applications. PLGA's biodegradability and biocompatibility allow for safe, controlled, and sustained release of drugs over time, which is particularly beneficial for chronic disorders that require long-term treatment. PLGA nanoparticles, in particular, offer enhanced stability for encapsulated drugs, protecting them from degradation before they reach their target site. The ability to modify the surface properties of PLGA nanoparticles further enables targeted drug delivery, enhancing the precision of treatment and improving patient outcomes (Danhier et al., 2012). As a result, PLGA continues to play a pivotal role in advancing drug delivery technologies, including those for cancer, infectious diseases, and tissue regeneration, CNS drug delivery overcoming the obstacles posed by the blood-brain barrier (BBB) (Costantino et al., 2006). The effective delivery of therapeutic ingredients to specific sites within the body in a controlled and sustained manner is essential to maximizing therapeutic efficacy while minimizing adverse side effects. PLGA-based drug delivery systems (DDS) offer a sophisticated solution to this challenge by enabling precise control over the administration of medications. By encapsulating drugs within biodegradable polymers, DDS can regulate drug release kinetics, extend therapeutic effects, and reduce undesired reactions.

However, designing and optimizing these systems present significant challenges due to the intricate interplay of physicochemical factors that govern drug release and thus having a good understanding of the underlying release mechanisms is paramount for the development of new DDS.

Traditionally, the development of DDS has heavily relied on experimental trial-and-error methodologies, which involve a meticulous exploration of various factors influencing drug release. These systems are engineered to regulate the release of medicinal substances over predetermined periods, enhancing treatment efficacy. However, the process requires extensive experimentation to determine the optimal composition and configuration of factors for specific diseases. This labor-intensive approach contributes to prolonged development timelines, complicating the creation of treatments for conditions ranging from cancer to chronic diseases. Furthermore, extracting meaningful insights from experimental data adds another layer of complexity (Fredenberg, et al., 2011)

To address these challenges, computational approaches, including mathematical modeling, molecular dynamics simulations, and machine learning (ML), have emerged as promising tools for accelerating DDS design and optimization. Mathematical models, while valuable for understanding drug release mechanisms and rates, are often material- and formulation-specific, limiting their generalizability and predictive power.

Additionally, combining multiple release mechanisms in a single model can be computationally intensive. Molecular dynamics simulations, though insightful at a molecular level, are computationally intensive and may not accurately capture real-world conditions (Versypt et al., 2013). In contrast, ML techniques leverage historical experimental data to identify key factors influencing drug release kinetics, offering an efficient and scalable solution for predictive modeling. By training ML models on comprehensive datasets, researchers can develop frameworks that guide the design and optimization of DDS, potentially transforming how these systems are developed (P. et al., 2023)

This study focuses on applying ML models to predict drug release kinetics and optimize dosing strategies for polymeric DDS, specifically those based on poly(lactide-co-glycolide) (PLGA). PLGA is a widely used biodegradable polymer with a well-established safety profile, though its compatibility with diverse drugs poses challenges. By training ML models on empirical drug release data, the study aims to provide insights into the acceleration of the development of PLGA-based DDS, optimize release profiles for specific diseases, and validate these findings through experimental testing. This integrated approach seeks to expedite drug development timelines and improve therapeutic outcomes for patients.

CHAPTER 2

LITERATURE REVIEW

2. LITERATURE REVIEW

Prior studies have extensively investigated various aspects of DDS to elucidate the drug release kinetics with numerous computational methodologies. To optimize the design of new DDS in a shorter timeframe and find important factors influencing the drug release from the polymerics systems to achieve controlled and sustained release, these methods have been utilized.

Traditional approaches, historically, DDS development relied heavily on empirical trial-and-error methods. These approaches, while effective to some extent, were time-consuming and resource-intensive, involving extensive experimentation to determine optimal drug-polymer compositions and formulations. This iterative process often resulted in prolonged development timelines with high failure rate, delaying the availability of new treatments.

To overcome the issues with the traditional approaches the Computational Strategies like– mathematical modeling has been employed to elucidate the release kinetics of various drug release mechanisms and assess the impact of formulation parameters (Versypt et al., 2013). Such models like– Zero-order kinetics where drug release occurs at a steady rate, independent of concentration, follows diffusion or erosion controlled mechanisms; First-order kinetics where release rate is proportional to the amount of drug that remains in the system, release rate decreases over time as the drug concentration in the DDS decreases, follows diffusion-controlled mechanisms; the Higuchi model (Higuchi, T., 1963) where drug is released from a matrix system, the release is diffusion-controlled and is proportional to the square root of time, indicating a decreasing rate of release over time as the drug near the surface of the matrix diffuses out; the Baker-Lonsdale model (Baker R.W. & Lonsdale H.S., 1974) is used for spherical matrix systems where the drug release is governed by diffusion from a spherical particle correcting the Higuchi model; the Hixson-Crowell model (Hixson & Crowell, 1931) addresses DDS where drug release is controlled by changes in the surface area and particle size of the drug formulation;

Korsmeyer-Peppas model (Korsmeyer R.W. et al., 1983), an empirical model that describes various release mechanisms (Fickian diffusion, non-Fickian transport, or swelling) depending on the diffusion exponent; the Lao model (Lao et al., 2008), a combined mathematical model that incorporates numerous drug release mechanisms such as burst release, diffusion, and degradation, as well as multiple biodegradable polymers such as PLGA and PCL, the model development process included forecasting the release kinetics of each material component in which PLGA model accounts for all three processes. While these many mathematical models provide valuable insights, their predictive accuracy is often constrained by the complexity of solute transport mechanisms and the specificity of models to materials, drugs, and formulations. There are no general mathematical models that can be employed to understand release mechanisms and predict release kinetics as most of them are release mechanisms specific, like diffusion-, swelling-, degradation-controlled (Fu et al., 2010).

Moreover, though molecular dynamics simulations offer detailed molecular-level insights into DDS behaviour, including interactions between drugs and polymers. But their high computational cost and limited real-world applicability pose significant challenges.

In recent years, machine learning (ML) (P. et al., 2023) has emerged as a transformative tool for predictive modeling in DDS design. By leveraging historical experimental data, ML models can identify critical factors influencing drug release kinetics and optimize formulation parameters with higher precision.

PLGA-based DDS have received significant attention due to the polymer's favourable properties, including biocompatibility, biodegradability, and versatility in drug delivery applications (Astete & Sabliov, 2006). ML techniques address several challenges inherent in traditional DDS development such as–Limited selection of biodegradable polymers safe for parenteral administration; Complex formulation optimization involving numerous interdependent variables; High failure rates associated with iterative experimental approaches. Despite advancements, significant work remains to enhance the predictive accuracy and reliability of ML models, particularly for PLGA-based systems. By refining these models, researchers can further accelerate development timelines, reduce resource consumption, and gain deeper insights into DDS behaviour.

This study builds on prior research by exploring the application of ML algorithms to predict fractional drug release from PLGA-based DDS and optimize their design. Through a comparative evaluation of multiple ML techniques, this work aims to contribute to the growing body of knowledge on computationally driven DDS development and pave the way for more efficient and effective therapeutic solutions

CHAPTER 3

AIMS & OBJECTIVES

3. AIMS AND OBJECTIVES

AIMS

To predict drug release kinetics of PLGA-based Drug Delivery Systems (DDS) using advanced machine learning techniques.

OBJECTIVES

- Train machine learning models to analyze historical empirical drug release data.
- Identify critical factors influencing the drug release profile of PLGA-based DDS.
- Enhance predictive accuracy to optimize the design of effective and controlled drug delivery systems.
- Explore the potential of data-driven approaches to overcome limitations of traditional mathematical modeling in drug release prediction.

CHAPTER 4

MATERIALS & METHODS

4. MATERIALS AND METHODS

To predict the release kinetics and find various important factors involved in drug release, the study here utilized the dataset collected from previously published paper (P. et al., 2023), specifically focusing on PLGA-based drug delivery systems. The used dataset contains a total of 1914 individual release profiles where 20 were unique drug-polymer combinations. And a total of 17 input physicochemical features of drugs, polymer, and their combinations are involved in all release profiles and these are– DP_Group(drug-polymer group, LA/GA(Lactic acid to glycolic acid ratio), Polymer_MW(molecular weight of polymer), CL Ratio(molar cross linking ratio),Drug_Tm(drug's melting temperature), Drug_Pka(partition coefficient), Initial D/M ratio(drug-to-polymer ratio), DLC(drug loading capacity), SA-V(surface area-to-volume ratio), SE(surfactant's percent in experiment),Drug_Mw(molecular weight of drugs), Drug_TPSA(topical polar surface area), Drug_NHA(number of heteroatoms), Drug_LogP(acid dissociation constant), Time(drug release points), T=0.25(drug release at 6hr), T=0.5(drug release at 12 hr), T=1.0(drug release at 24 hr) and output Release(fractional drug release).

4.1 Data Preprocessing

To make the data suitable for training machine learning (ML) models, the following steps had been taken: Feature selection and Feature engineering: In order to train the model, all 17 molecular and physicochemical descriptors were first employed as input features. To avoid multicollinearity and enhance the generalizability of the model, related features with Spearman's rank correlation values greater than 0.9 were eliminated. Then, to further refine the best model and determine the significance of the features, the seven most top features of the best model out of the 17 input feature models were chosen; Feature Scaling: Standardization (Z-score normalization) was used to scale the input features to guarantee that each feature contributed equally to the model. This procedure ensures that all variables are within a range by transforming features to have a mean of 0 and a standard deviation of 1, especially for SVR.

4.2 Machine Learning Models

In this study, eight different regression models were implemented to predict fractional drug release. These models include:

1. **Linear Regression**
   A simple model that assumes a linear relationship between the features and the target variable. It is used as a baseline for comparison.
2. **K-Neighbors Regressor**
   A non-parametric method that predicts the target value based on the average of the target values of the k-nearest neighbors in the feature space.
3. **Support Vector Regressor (SVR)**
   This model uses the support vector machine (SVM) algorithm to find a hyperplane that best fits the data in a high-dimensional space. It is effective in handling non-linear relationships by using kernel functions.
4. **Decision Tree Regressor**
   A non-linear model that splits the feature space into regions based on feature values, minimizing variance within each region. It recursively splits the dataset at each node to form a tree structure.
5. **Random Forest Regressor**
   An ensemble method consisting of multiple decision trees. Each tree is trained on a random subset of the data, and the final prediction is made by averaging the predictions of all trees. This reduces overfitting and variance, enhancing the model's accuracy.
6. **AdaBoost Regressor**
   A boosting method that combines multiple weak learners (typically decision trees) to create a strong learner. Each new model focuses on correcting the errors of the previous ones, improving overall prediction accuracy.
7. **Gradient Boosting Regressor**
   Another boosting technique that builds an ensemble of models in a sequential manner. Each new model corrects the residual errors of the previously trained models. This model works well with smaller datasets by iteratively refining predictions.
8. **XGBoost Regressor**
   An optimized version of gradient boosting, XGBoost includes regularization to avoid overfitting and can handle large datasets efficiently. It is known for its

scalability and high performance in regression tasks.

4.3 Data Splitting for model training

To ensure a robust evaluation of the machine learning models, the dataset was divided into training, validation, and testing subsets using a combination of GroupShuffleSplit and Group-KFold cross validation strategies from Scikit-learn library of python (Pedregosa et al., 2011). These methods ensure that data from the same drug-polymer combination group is not split across training, validation, or testing sets, thus preventing data leakage and maintaining the integrity of model evaluation. a) Test Set: The dataset was grouped by the drug-polymer combinations (DP_Group), and 20% of these groups were randomly held out as the test set using GroupShuffleSplit. This test set is used exclusively for evaluating the model's performance on unseen drug-polymer combinations. Additionally, b) Training and Validation Sets: The remaining 80% of the data was used for training and validation. Within this subset, GroupKFold cross-validation(k=10) was applied to divide the data into 10 folds. And for each fold, 90% of the training data from the current fold was used to train the model while 10% of the training data was reserved for validation to fine-tune hyperparameters.

4.4 Model selection and Hyperparameter tuning

The best performing Random Forest regressor was chosen for further analysis or optimizations. The model's performance was optimized using a GridSearchCV approach from Scikit-learn library of python(Pedregosa et al., 2011)to systematically explore the best hyperparameters combination. The grid search was conducted within each training fold, with the negative mean absolute error (MAE) used as the scoring metric. The best-performing fold's hyperparameters were identified and used to train the model on that fold's training data.

4.5 Model Training on optimized hyperparameters and Evaluation

For each fold in the GroupKFold, the model was trained using the best hyperparameters identified by GridSearchCV. The trained model was evaluated on the validation set, and the absolute errors between predicted and actual values were recorded for each fold. For the best model selection, the model with the lowest validation error (highest negative MAE score) across all folds was selected

as the best model.

This model was retrained on the entire training set (80% of the dataset) to ensure it benefited from all available training data. The best model was used to predict the target variable (Fractional drug release) for the test set. The absolute errors between predicted and actual values were calculated for the test set to evaluate the model's generalization performance.

4.6 Performance Metrics

The following metrics were calculated to assess the models' performance:

1. **MAE (Mean Absolute Error)**: The average magnitude of the absolute differences between estimated and actual values.

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

2. **MSE** (Mean Squared Error)-It averages the square of the difference between actual value and estimated value

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

3. **R^2 (**Coefficient of determination) regression score function to evaluate how well the regression line fits the actual data. it ranges from 0 to 1

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

CHAPTER 5

RESULTS AND DISCUSSIONS

5. RESULTS AND DISCUSSIONS

The data contains both cancerous as well as non-cancerous drugs release profiles with almost 53% for cancerous and 47% for non-cancerous (Fig. 1) with 20 unique drugs (Table–1). Out of 17 features that were taken into consideration for drug release kinetics, many of them found out to be correlated (using spearman's correlation) (Fig.2).

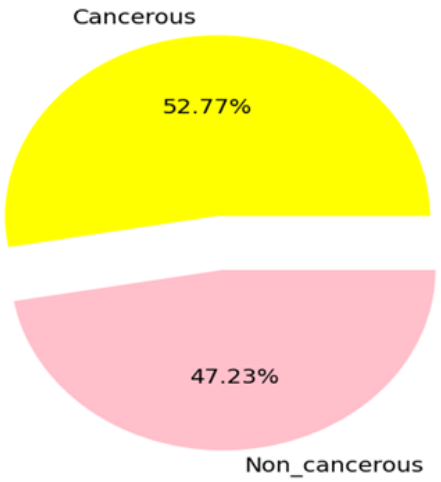

Fig.1: Total data distribution in terms of cancerous and non-cancerous

DP_Group(DDS)	Drug	Category
5-FU-PLGA	5-Fluorouracil	Cancerous
CAF-PLGA	Caffeine	Non-Cancerous
CBD-PLGA	Cannabidiol	Non-Cancerous
DEX-PLGA	Dexamethasone	Non-Cancerous
DPP-PLGA	DPP	Non-Cancerous
GEF-PLGA	Gefitinib	Cancerous
HPA-PLGA	Hydroxyapatite	Non-Cancerous
IBP-PLGA	Ibuprofen	Non-Cancerous
LDC-PLGA	Lidocaine	Non-Cancerous
LPA-PLGA	Lipoic acid	Non-Cancerous
LTZ-PLGA	Letrozole	Cancerous
PRC-PLGA	Pracinostat	Cancerous
PTX-PLGA	Paclitaxel	Cancerous
PTX-PLGA-co-PALA	Paclitaxel	Cancerous
PTX-PLGA-co-PAVL	Paclitaxel	Cancerous
TAA-PLGA	Triamcinolone acetonide	Cancerous
TAA-PLGA-co-PALA	TAA	Non-Cancerous
TAA-PLGA-co-PAVL	TAA	Non-Cancerous
TMZ-PLGA	Temozolomide	Cancerous
TTD-PLGA	TAA	Non-Cancerous

Table 1: All the unique PLGA-based DDS with drug names and categories.

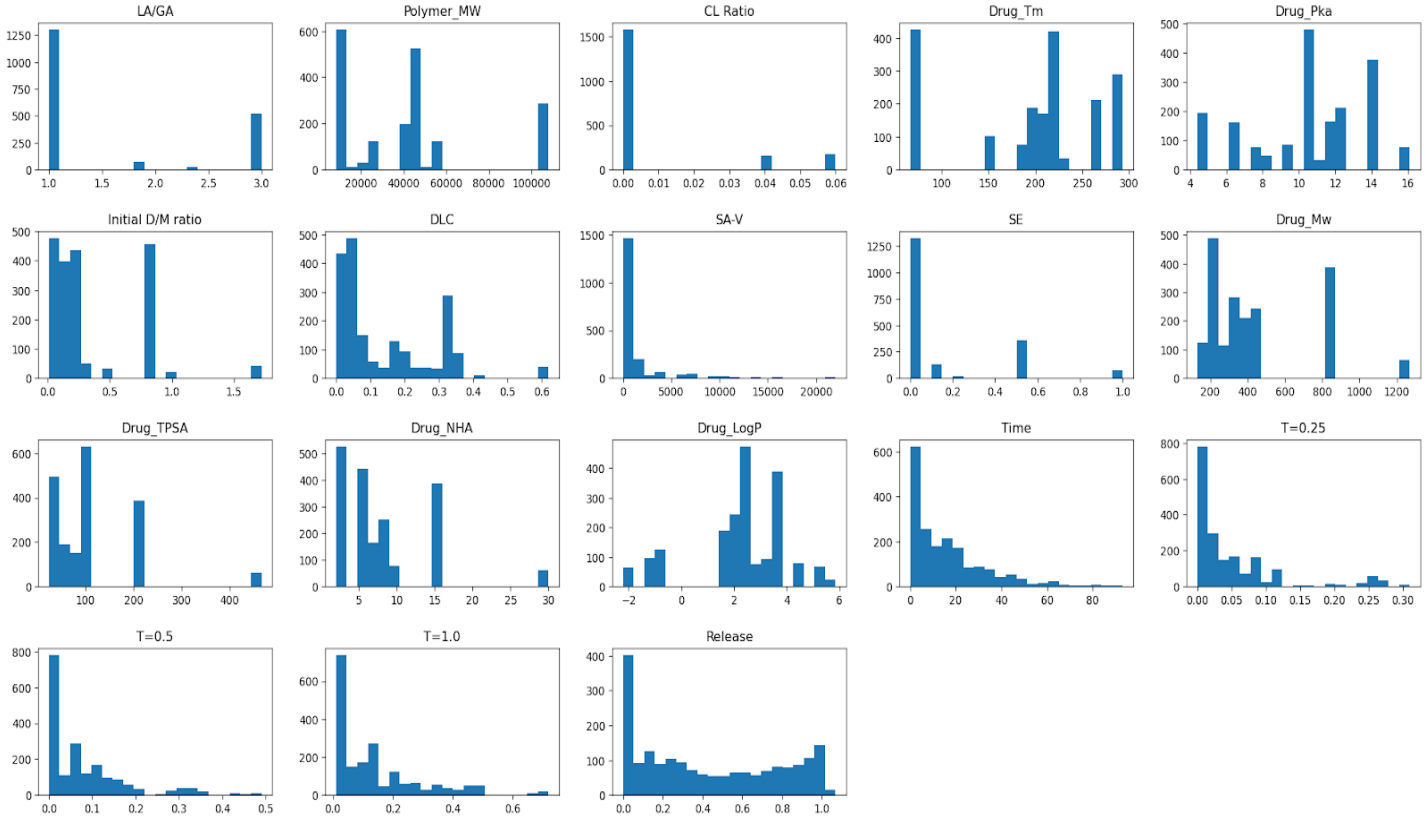

Fig. 2: All the input features' distribution that's been utilized to train ml models for the prediction of release kinetics. Like in most release profiles, LA/GA ratio is 1 and 3; CL ratio is 0 in almost all the profiles; SA-V remains in the region of 0–2500 etc.

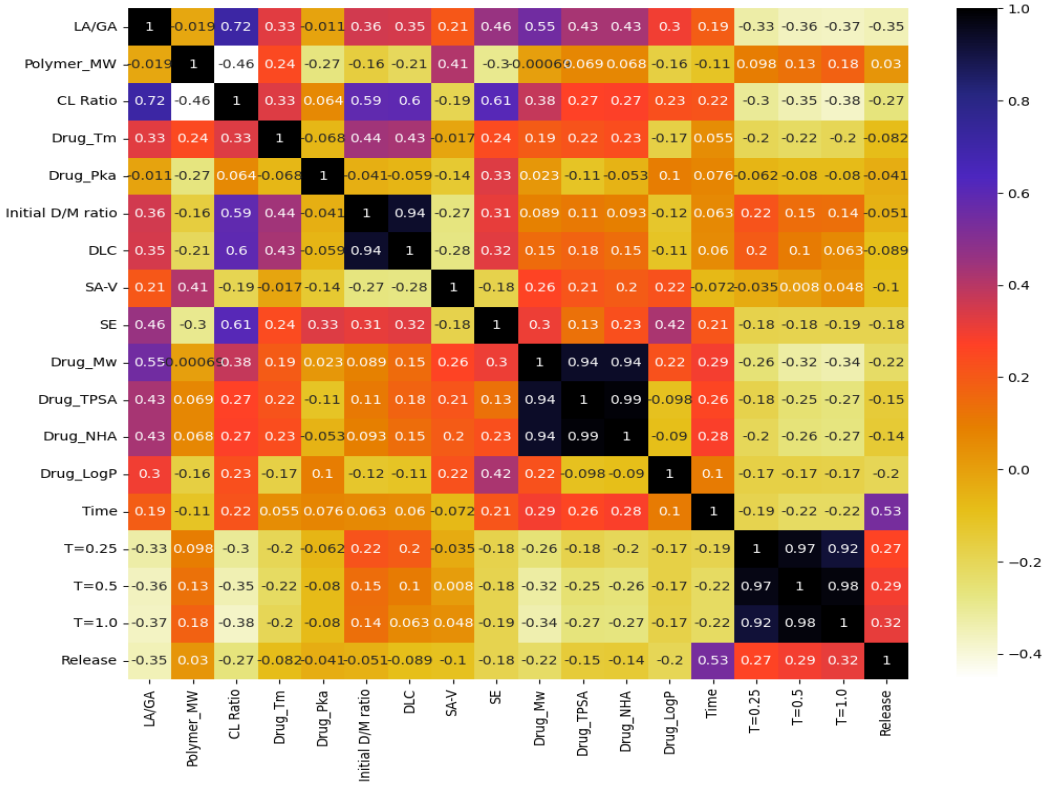

Fig. 3: Pearson correlation heatmap matrix and Features where correlation is more that 0.90 – between (DLC, Initial D/M ratio), among (Drug_NHA, Drug_Mw Drug_TPSA), among (T=0.5, T=1.0, T=0.25)

5.1 Comparative analysis of all the trained models

The performance of various machine learning models (in total 8) was evaluated using three sets of input features: 17 features (all input features), 12 uncorrelated features (after removing highly correlated ones where pearson's correlation were more than 0.90(Fig.1)), and 7 selected features (selected based on feature importance from RF model where importance is more than 0.01). Optimization techniques were applied to enhance the predictive accuracy of each model. The results are summarized below:

5.2 Performance metrics on 17 input features and all the models' prediction on test DDS

The Random Forest (RF) and XGBoost, two ensemble tree-based models, outperformed all other models evaluated (Table−2), which is consistent with previous findings (Grinsztajn et al., 2022) that tree-based models are effective for medium size data. The two best models were then investigated for further development, and it was observed that the RF model performed marginally better with modified optimized tuned hyperparameters, with the MAE of RF(testing) being 0.1144 and of XGBoost(testing) being 0.1388(Table−3).

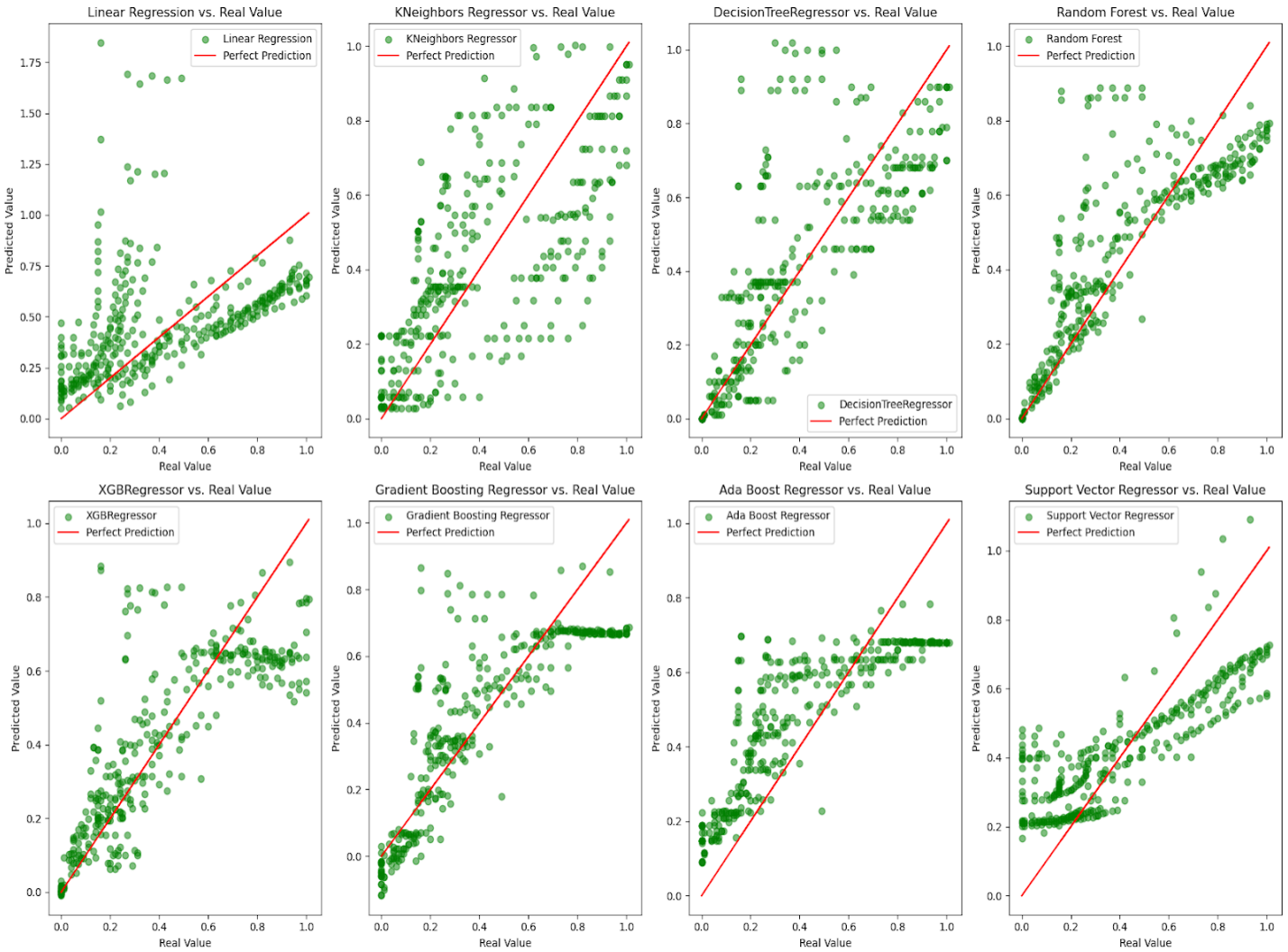

Fig.4: All the models' prediction on test DDS vs experimental drug release with red line representing perfect prediction, and the Random Forest model's prediction points are closer to perfect fit than other model's prediction

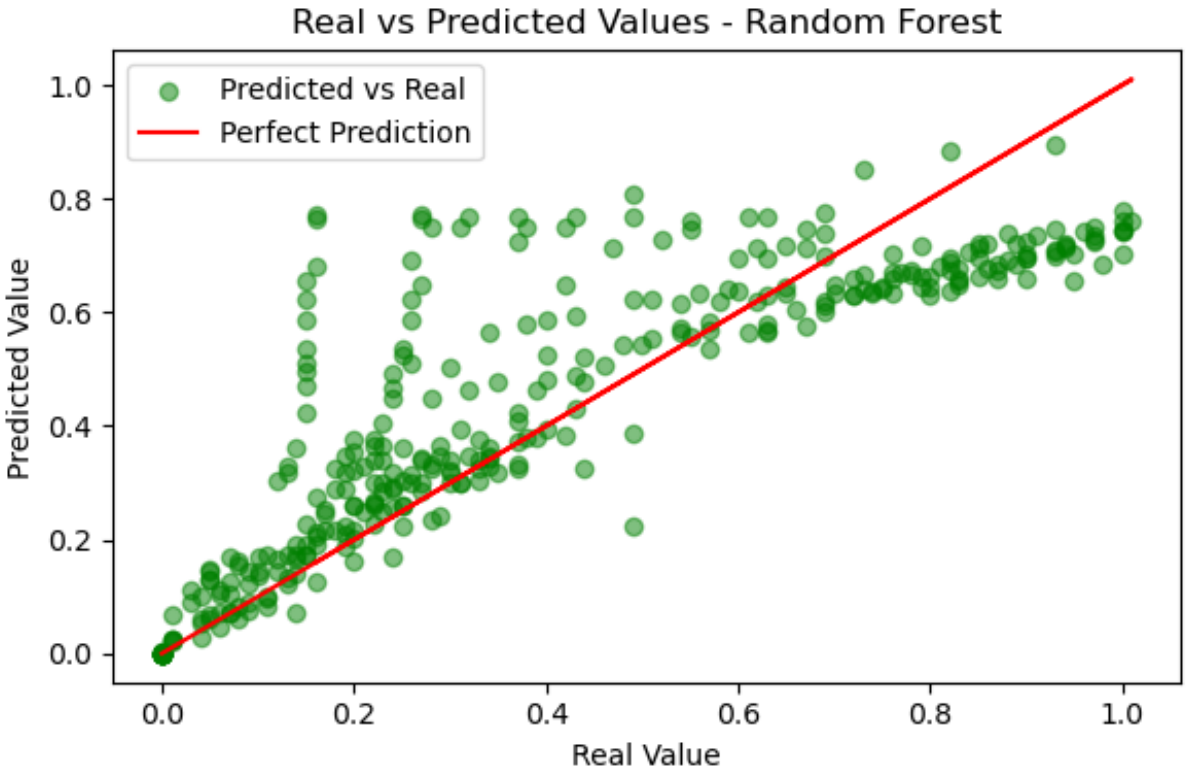

Fig.5: RF models' prediction on test DDS vs experimental drug release

5.3 Performance on Reduced Input Features

On 12 Uncorrelated Features (Table–4): Random Forest maintained strong performance with a Testing R^2 = 0.6865, comparable to the full 17 features. This indicates that removing highly correlated features did not adversely affect the model's predictive capabilities. And with **7** top selected features (Table–5): Random Forest's performance improved slightly, achieving a Testing R^2 = 0.7101, which was higher than the performance with 17 features, which is also visible from Fig.4 & Fig.5 wherein RF model's most predictions were close to perfect line(red) compared to other models' predictions. The testing metrics for this setup included: MSE=0.0264, MAE=0.1135, which is consistent with previous findings (P. et al., 2023). This improvement highlights the benefits of feature selection in enhancing model generalizability and it tells us that we can achieve the same or better results using only 7 features that are physicochemically important.

Models	MSE (Training)	MSE (Testing)	MAE (Training)	MAE (Testing)	R2 (Training)	R2 (Testing)
Linear Regression	0.0403	0.1057	0.1609	0.2332	0.6854	-0.1620
K-Neighbors	0.0067	0.1045	0.0480	0.2605	0.9478	-0.1480
SVR	0.1236	0.1964	0.2965	0.3288	0.0348	-1.1579
Decision Tree	0.0001	0.0566	0.0013	0.1710	0.9989	0.3783
AdaBoost	0.0273	0.0356	0.1451	0.1601	0.7869	0.6083
GradientBoost	0.0060	0.0312	0.0566	0.1284	0.9533	0.6570
XG-Boost	0.0003	0.0305	0.0075	0.1216	0.9980	0.6644
Random Forest	0.0006	0.0297	0.0136	0.1172	0.9956	0.6738

Table–2: Models performance on all 17 input features before Hyperparameter tuning

Models	MSE (Training)	MSE (Testing)	MAE (Training)	MAE (Testing)	R2 (Training)	R2 (Testing)
XGBoost	0.0010	0.0383	0.0213	0.1388	0.9922	0.5793
Random Forest	0.0006	0.0285	0.0141	0.1144	0.9952	0.6865

Table-3: Best tree-based models' performance after Hyperparameter tuning

Random Forest	MSE	MAE	R2	STD
Training	0.0006	0.0141	0.9952	0.0204
Testing	0.0285	0.1144	0.6865	0.1243

Table–4: Random Forest on 12 Un-correlated features

Random Forest	MSE	MAE	R2	STD
Training	0.0007	0.0148	0.9949	0.0211
Testing	0.0264	0.1135	0.7101	0.1169

Table–5: Random Forest on 7 most significant features

5.4 Discussion

5.4.1 Model Performance Across Feature Sets

The Random Forest model consistently outperformed all other models across the different feature sets, showcasing its robustness and ability to handle complex sparse datasets effectively. This was evident from its high R^2 scores and low error metrics (MSE and MAE). Ensemble methods like Gradient Boost and XGBoost also showed strong performance, though they lagged slightly behind Random Forest. These models excel in capturing intricate patterns in the data, making them suitable for predictive tasks in biopolymer-based drug delivery systems. In contrast, simpler models such as Linear Regression and SVR struggled significantly. The poor performance of these models indicates that the dataset exhibits strong nonlinearity and interactions among features that cannot be effectively captured by linear models or traditional regression approaches.

5.4.2 Significance of Feature Selection and Interpretation of the RF model

Feature selection played a critical role in improving the performance and interpretability of the models. Random Forest's feature importance function analysis of the 17, 12, and 7 features was used to assess the significance of the different input features in producing such fractional drug release forecasts. Their importance underscores their significant influence on the underlying processes of drug release in PLGA-based drug delivery systems. Removing highly correlated features (e.g., reducing to 12 uncorrelated features) and focusing on the most relevant ones and informative features (7 selected features) helped mitigate overfitting and noise in the data with improved generalization on the test dataset. The input features are sorted from top to bottom in (Fig. 6 & Fig.7) according to their importance/influence on the model's output. Time had the most influence on the fractional drug release forecasts, as seen by its listing in the first row. The 7 selected features (Polymer_MW, Drug_Mw, CL Ratio, SA-V, DLC, Time, and T=1) were found to be critical in driving the model's predictions. Interestingly, descriptors of the drug and polymer's physicochemical characteristics (i.e., molecular weight of drugs and polymer) are among the most significant attributes. This suggests that among the DDS's constituents, the RF model has identified that the drug's and the polymer's molecular weights have one of the most significant effects on fractional drug release. The remaining physicochemical and molecular parameters seem to also contribute to the model's fractional drug release prediction. (Fig.6, Fig.7 & Fig.8 do not take into consideration the possibility of synergy between input features; rather, it just illustrates the impact of each individual element.)

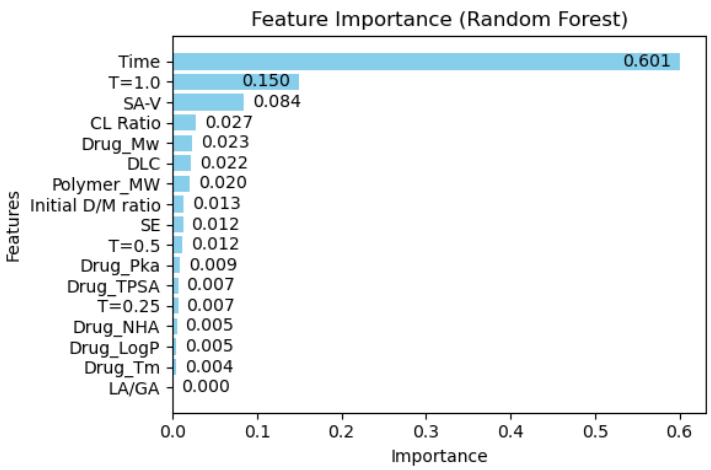

Fig.6 Feature importance in Random Forest for 17 input features

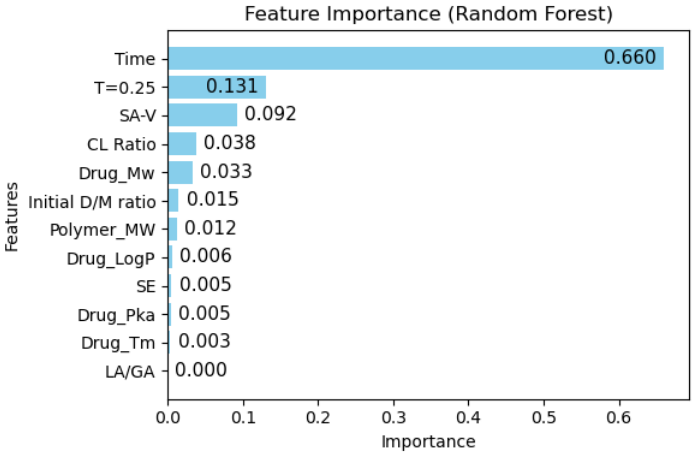

Fig.7 Feature importance in Random Forest for 12 un-correlated features.

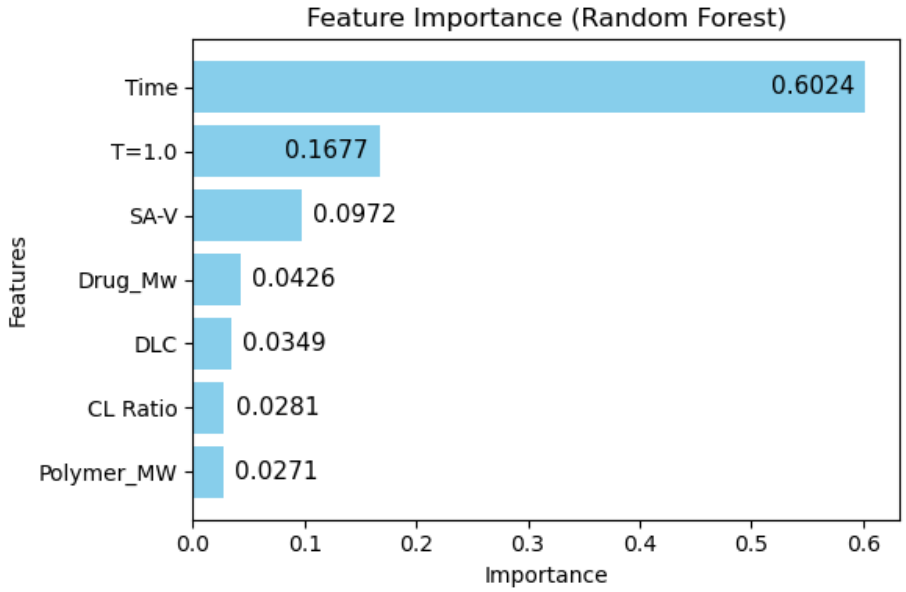

Fig.8 Feature importance in Random Forest for 7 input features

5.4.3 Comparison of all the models

The ensemble models (Random Forest, Gradient Boost, and XGBoost) leveraged their ability to combine multiple weak learners to achieve high accuracy and robustness. Their performance reflects their capacity to handle complex feature interactions and nonlinearity in the dataset. The poor performance of SVR, with a negative R^2 score (-1.1579), suggests that it struggled with the data's complexity, potentially due to its sensitivity to hyperparameter tuning and inability to handle high-dimensional, nonlinear datasets without extensive preprocessing. Linear Regression's negative R^2 score further illustrates the limitations of linear models for this type of problem, where relationships between variables are not strictly linear.

5.4.4 RF model's prediction vs experimental result of Individual test DDS and limitations

The test set (5-FU-PLGA, CBD-PLGA, PTX-PLGA) drug–polymer combinations are displayed against their corresponding experimental drug release profiles to demonstrate the estimated fractional drug release profiles produced by the trained Random Forest model (7-features model). Overall, the predicted and experimental drug release patterns coincided rather well with some limitations. We can see in all subplots, the predicted (red) and actual (blue) drug release profiles show good agreement during the initial phase of drug release. In case of 5-FU-PLGA, for Exp Index = 84: Both curves match well up to approximately 5 days, with ~50% initial drug release. And, after 5 days, the predicted release begins to plateau around 60–65%, while the actual release continues to increase steadily towards 100%. For the Exp Index = 85: Both curves match well up to approximately 3 days, with ~ 62% initial drug release. After 3 days, the predicted release plateaus around 65–70%, while the actual release continues to increase gradually towards 100%(Fig.9)

Additionally, in case of CBD-PLGA– For Exp Index = 36: Both the actual and predicted release curves match closely up to ~10 days. After 10 days, the predicted release progresses faster, reaching a plateau (~85%) much earlier than the actual release curve. The actual release continues to increase steadily and reaches nearly 90% by 40 days. The predicted release plateaus earlier and does not capture the sustained release phase as accurately.

And for Exp Index = 37: Both curves show good agreement up to ~8 days. After 8 days, the predicted release slows down and plateaus around 70%, while the actual release continues to increase steadily, reaching close to 80% at the end of 40 days (Fig.10).

Also, for PTX-PLGA, Exp Index = 50: Both the actual and predicted release curves match closely up to ~10 days. After 10 days, the predicted release progresses faster, reaching a plateau at ~75% much earlier than the actual release curve. The actual release continues to increase steadily and reaches nearly ~75% by 30 days. The predicted release plateaus earlier and does not capture the sustained release phase as accurately. For PTX-PLGA, Exp Index = 51: Both the actual and predicted release curves show good agreement up to ~15 days. After 15 days, the predicted release progresses faster and plateaus around ~80%, while the actual release continues to increase steadily, reaching close to ~70% at the end of 30 days. The predicted release overestimates the late-stage cumulative release and does not accurately reflect the slower sustained release observed in the actual data (Fig.11).

Thus, all experimental indices reveal that the predictive model performs well during the early phase (burst release) but struggles to capture the long-term, especially the late phase release behaviour of the drug. This pattern highlights the model's tendency to overpredict the release fraction during the mid-to-late, emphasizing the need for better representation of diffusion- and degradation-controlled mechanisms.

7-Input feature Random Forest model's prediction:

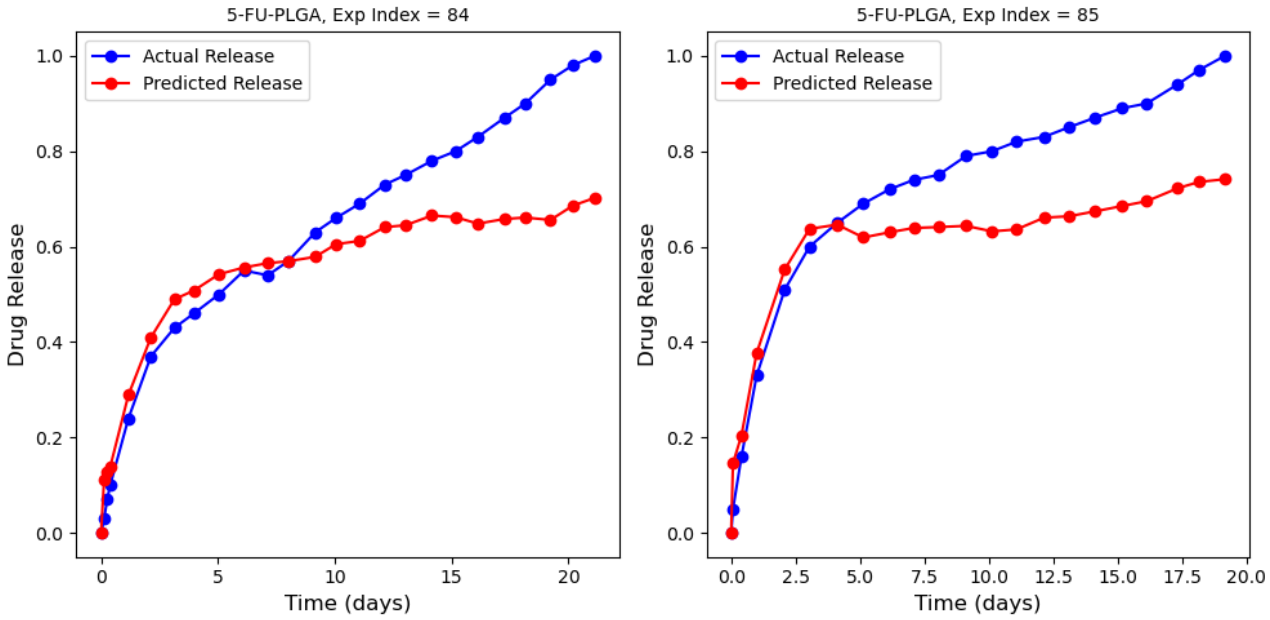

Fig.9 Experimental and predicted drug(5-Fluorouracil) release over time in 5-FU-PLGA DDS In index-84, the model can predict the initial ~60% drug release correctly and in index-85, it is able to predict almost ~65% drug release accurately.

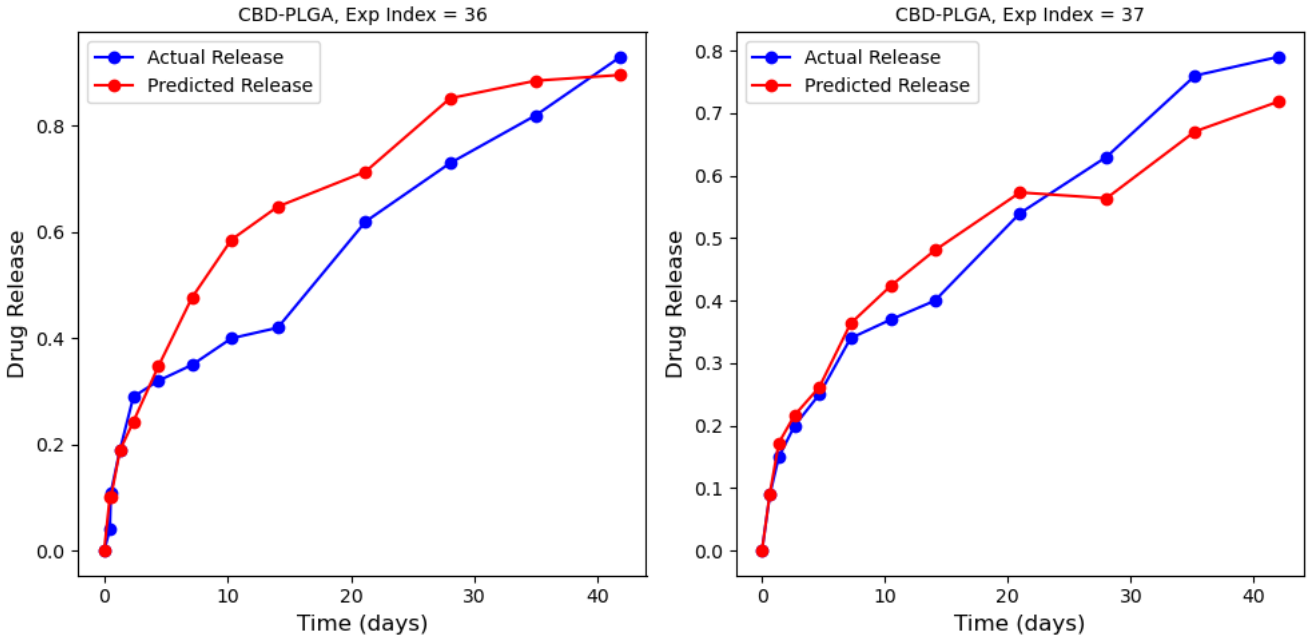

Fig.10: Experimental and predicted drug (Cannabidiol) release over time in CBD-PLGA DDS Here, in both the indices it's clear that the model can predict the initial burst release accurately while maintaining the same proportion with experimental releases in the later stage.

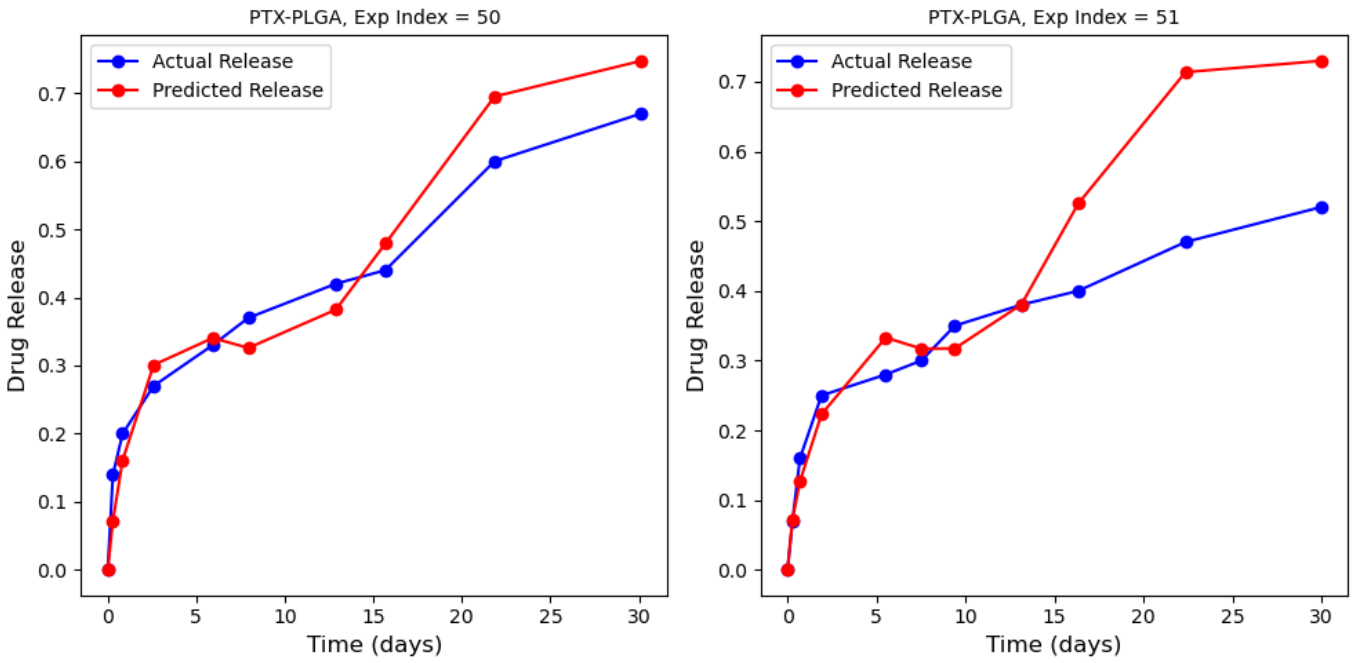

Fig.11: Experimental and predicted drug (Paclitaxel) release over time in PTX-PLGA DDS Here, in both the indices the model can predict the initial burst and a bit of sustained release accurately while maintaining the same proportion with experimental releases in the later stage.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

6. CONCLUSIONS & FUTURE WORK

The results of this study provide significant insights into the predictive modeling of biopolymer-based drug delivery systems where there is sparse and medium size data, especially for PLGA-based DDS. Feature selection methods and ensemble learning algorithms like Random Forest can provide accurate predictions, aiding in the design and optimization of drug delivery systems. Identifying key features that drive drug release kinetics (e.g. molecular weight of drugs and polymer, drug loading capacity, surface area-to-volume ratios, cross-linking ratios) can guide experimental efforts, reducing the need for exhaustive trial-and-error methods. While Random Forest showed excellent performance in predicting the initial drug release (burst release), further exploration of hyperparameter tuning and advanced feature engineering techniques with more data could improve results and make models more accurate for late phase drug delivery prediction. Integrating domain knowledge into feature selection and model design may enhance the interpretability and applicability of predictive models in real-world scenarios. Future work could also focus on incorporating additional features or external datasets to validate the generalizability of the models across diverse drug delivery systems.

REFERENCES

1. Arifin, D. Y., Lee, L. Y., & Wang, C. H. (2006). Mathematical modeling and simulation of drug release from microspheres: Implications to drug delivery systems. *Advanced Drug Delivery Reviews, 58*(12–13), 1274–1325. https://doi.org/10.1016/j.addr.2006.09.007

2. Baker, R. W., & Lonsdale, H. S. (1974). Controlled release of biologically active agents. *Plenum Press*.

3. Bannigan, P., Bao, Z., Hickman, R. J., Aldeghi, M., Häse, F., Aspuru-Guzik, A., & Allen, C. (2023). Machine learning models to accelerate the design of polymeric long-acting injectables. *Nature Communications, 14*(1), 35. https://doi.org/10.1038/s41467-022-35343-w

4. Casalini, T., Rossi, F., Lazzari, S., Perale, G., & Masi, M. (2014). Mathematical modeling of PLGA microparticles: From polymer degradation to drug release. *Molecular Pharmaceutics, 11*(11), 4036–4048. https://doi.org/10.1021/mp500078u

5. Chang, S. (Ed.). (2003). *Encyclopedia of biopharmaceutical statistics*. Informa Health Care.

6. Ford Versypt, A. N., Pack, D. W., & Braatz, R. D. (2013). Mathematical modeling of drug delivery from autocatalytically degradable PLGA microspheres: A review. *Journal of Controlled Release, 165*(1), 29–37. https://doi.org/10.1016/j.jconrel.2012.10.015

7. Fredenberg, S., Wahlgren, M., Reslow, M., & Axelsson, A. (2011). The mechanisms of drug release in poly(lactic-co-glycolic acid)-based drug delivery systems: A review. *International Journal of Pharmaceutics, 415*(1–2), 34–52. https://doi.org/10.1016/j.ijpharm.2011.05.049

8. Higuchi, T. (1963). Mechanism of sustained-action medication: Theoretical analysis of rate of release of solid drugs dispersed in solid matrices. *Journal of Pharmaceutical Sciences, 52*, 1145–1149. https://doi.org/10.1002/jps.2600521210

9. Hopfenberg, H. B. (1976). Controlled release polymeric formulations. *American Chemical Society*.

10. Korsmeyer, R. W., Peppas, N. A., Gurny, R., Doelker, E., & Buri, P. (1983). Mechanisms of solute release from porous hydrophilic polymers. *International Journal of Pharmaceutics, 15*(1), 25–35. https://doi.org/10.1016/0378-5173(83)90064-9

11. Leong, K. W., & Langer, R. (1987). Polymeric controlled drug delivery. *Advanced Drug Delivery Reviews, 1*, 199–233.

12. Lao, L. L., Venkatraman, S. S., & Peppas, N. A. (2008). Modeling of drug release from biodegradable polymer blends. *European Journal of Pharmaceutics and Biopharmaceutics, 70*(3), 796–803. https://doi.org/10.1016/j.ejpb.2008.05.024

13. Hixson, A. W., & Crowell, J. H. (1931). *Ind. Eng. Chem.*, *23*, 923.

14. Peppas, N. A., & Sahlin, J. J. (1989). A simple equation for the description of solute release. III. Coupling of diffusion and relaxation. *International Journal of Pharmaceutics, 57*, 169–172.

15. Adepu, S., & Ramakrishna, S. (2021). Controlled drug delivery systems: Current status and future directions. *Molecules, 26*(19), 5905. https://doi.org/10.3390/molecules26195905

16. Astete, C. E., & Sabliov, C. M. (2006). Synthesis and characterization of PLGA nanoparticles. Journal of biomaterials science. Polymer edition, 17(3), 247–289. https://doi.org/10.1163/156856206775997322

17. Danhier, F., Ansorena, E., Silva, J. M., Coco, R., Le Breton, A., & Préat, V. (2012). PLGA-based nanoparticles: an overview of biomedical applications. Journal of controlled release : official journal of the Controlled Release Society, 161(2), 505–522. https://doi.org/10.1016/j.jconrel.2012.01.043

18. Costantino, L., Gandolfi, F., Bossy-Nobs, L., Tosi, G., Gurny, R., Rivasi, F., Vandelli, M. A., & Forni, F. (2006). Nanoparticulate drug carriers based on hybrid poly(D,L-lactide-co-glycolide)-dendron structures. Biomaterials, 27(26), 4635–4645. https://doi.org/10.1016/j.biomaterials.2006.04.026

19. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

20. Léo Grinsztajn, Edouard Oyallon, Gaël Varoquaux. Why do tree-based models still outperform deep learning on tabular data? NeurIPS 2022 Datasets and Benchmarks (2022).