CS 4850 – Spring 2025

AT-1 —Nano Drug Delivery Efficiency Prediction using Machine Learning

Zachary Mitchell

Professor Perry

April 28, 2025

Website

Github

STATS AND STATUS		
LOC	474	
Components/Tools	Google Collab	
Hours Estimate	300	
Hours Actual	300	
Status	Project is 100% complete and working as designed	

Nano Drug Delivery Efficiency Prediction using Machine Learning

Zachary Mitchell[#1]
College of Computing and Software, Kennesaw State University
1100 South Marietta Pkwy SE, Marietta, Georgia
zmitch18@student.kennesaw.edu

Abstract - This research explores the use of ensemble learning and synthetic data generation techniques to improve the prediction of delivery efficiency in targeted nano drug delivery systems. Specifically, our objective is to estimate delivery efficiency to tumors by training ensemble models on a combination of real and synthetically generated datasets. While previous research on this topic heavily relies on individual predictive models, our research investigates whether stacking ensemble approaches can enhance the predictive accuracy of drug efficiency. A combination of 3 to 10 weak learners were used as the base models and various meta models were tested to see where the accuracy was the highest. Our findings suggest that ensemble models, specifically those that implement bagging strategies and Support Vector Regression (SVR) show potential in improving delivery efficiency predictions. Overall this research contributes to the growing field of ML assisted nanomedicine by offering a unique approach to enhancing drug delivery efficiency.

Keywords — Ensemble Learning, Stacked Ensemble Models, Nanomedicine, Targeted Drug Delivery, Machine Learning in Healthcare, Drug Delivery Prediction, Regression Models, Synthetic Data Generation, Generative Adversarial Networks (GANs), Synthetic Data Vault (SDV), Nanoparticles, Biomedical Machine Learning

I. Introduction

In recent years, the field of medicine has made much progress in technology, biology and engineering. Despite these developments, medical problems such as cancer remain without a solution. Traditional healthcare approaches often face limitations due to the use of systemic administration, a method in which medicine is delivered throughout the entire body via the bloodstream without an intended target. While this method ensures that the drugs reach various parts of the body, both the diseased and healthy cells are exposed to the drug. As a result these treatments can lead to toxic side effects and damage to healthy tissue. For example in chemotherapy, a common treatment for cancer, affects not only the cancerous cells but also the healthy cells leading to side effects such as hair loss and immune suppression. In some cases an increased dosage of the drug is required to be delivered to the patient because of the low precision of the delivery of the drug.

To address these concerns, nanomedicine has emerged as a promising solution. Nanomedicine is the use of nanoparticles such as polymers and metal oxides that can be applied in medical applications, such as targeted drug delivery. In targeted drug delivery the drugs are encapsulated in nanoparticles or liposomes and engineered to reach specific disease sites in the body using biological markers or external stimuli such as pH, heat, or magnetism. This approach enhances drug delivery precision by ensuring that the medication accumulates solely at the disease site while minimizing exposure to healthy cells. Overall targeted drug delivery has the potential to significantly improve accuracy of medication delivery in complex diseases like cancer.

However designing effective personalized nanoparticle formulations for each patient is a complex task. Everyone is different and has a vast number of characteristics that have to be considered such as size, shape, surface charge, material type, and targeting mechanism. All of these features create a high-dimensional problem. The relationship between all these features are complex to understand and most likely nonlinear. Manually exploring every combination of features is time consuming and expensive. This is where machine learning (ML) comes in.

ML offers an efficient way to predict the delivery efficiency of nanoparticles by learning patterns from existing data. By modeling the complex, nonlinear relationships between nanoparticle properties and delivery efficiency, ML models can guide the design of more effective formulations. In particular ensemble techniques which combine the strength

of multiple individual models called weak/ base models to improve the performance in regression tasks.

In this research I explored the applications of ensemble models to improve the prediction of targeted drug delivery of tumors..

Ⅱ. Related Works

Ensemble Learning has become a major strategy in regression tasks such as improving prediction accuracy for complex biology problems. In [1] various ensemble learning strategies including bagging, tracking and boosting were compared for drug-target interaction prediction. This study displayed that random forest ensemble models significantly outperform singular models by reducing bias and variance. The research also emphasized that stacking can be a powerful technique when the models are selecting carefully. With the right model combinations issues like overfitting and high computational costs can be avoided.

Ensemble learning has become a prominent strategy in improving prediction accuracy for complex biological problems such as drug delivery and drug-target interaction. In [1], different ensemble learning strategies, including bagging, stacking, and boosting, were compared for drug-target interaction (DTI) prediction. The study highlighted that ensemble methods, particularly Random Forest and Extra Trees, significantly outperform individual models by reducing bias and variance. It also emphasized that stacking methods, though powerful, require careful model selection and tuning to avoid issues like overfitting and high computational costs. Furthermore, bagging methods such as Random Forest demonstrated strong predictive capabilities and were less prone to overfitting, making them favorable for biomedical datasets where sample size may be limited.

Similarly [2] implemented multiple machine learning models, including linear regression, random forest, support and deep neural networks (DNN) to predict the delivery efficiency (DE) of nanoparticles in tumors and major organs based on physicochemical properties. This study found that DNN models generally produced the best performance, outperforming all the traditional ML models. DNN achieved a r-squared .41 for tumor delivery efficiency prediction which is the highest achieved within their research. Feature importance analysis later revealed that the material composition and cancer type played important roles in influencing delivery outcomes. Importantly their research displayed the potential of ML models in improving drug delivery predictions.

.

Ⅲ. Methodology

The dataset used in this research comes from the Journal of Controlled Release and consists of 423 rows of data pertaining to nanoparticle delivery to tumors. This dataset includes features like type, Size, Shape, Zeta Potential, and Administration Route, with the target variable being the delivery efficiency to the tumor. During preprocessing the data went through missing value handling, removing null target variable, outlier detection, one hot encoding and normalization to ensure the data is prepared for modeling. After cleaning, the final dataset contained 403 rows. The data was then split into training and testing sets to evaluate model performance.

A stacked ensemble architecture was implemented for the predictive modeling framework. Bagged Random Forest Regressors (RFs) were used as base models and trained on bootstrapped subsets of the data to introduce variance among learners. Initially, five base RF models were trained and based on the results I continued to increase/decrease the number of RFs until I stopped seeing an increase in accuracy. Predictions from the base models were then passed to various meta-models including Linear Regression, Ridge Regression (RidgeCV), Random Forest Regressor, XGBoost Regressor, LightGBM Regressor, and Support Vector Regressor (SVR).

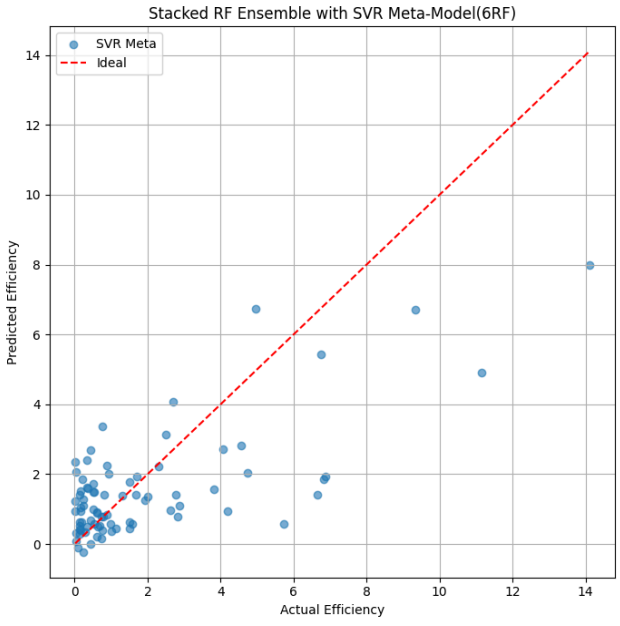
Each ensemble model was evaluated based on two primary metrics: R^2 Score and Root Mean Squared Error (RMSE). R^2 score is a coefficient of determination used to assess how well the variance in the target is captured while RSME is a measurement that helps understand how much difference there is between the value predicted by the model and the actual value. Scatter plots of the actual and predicted delivery efficiencies were also generated to visually assess the predictions

To optimize performance, experiments were conducted changing the number of base models from a range of 3 to 5 RF's and pruning weaker base models when necessary. In the final phase of optimizationHHyperparameter tuning was applied to the top performing model

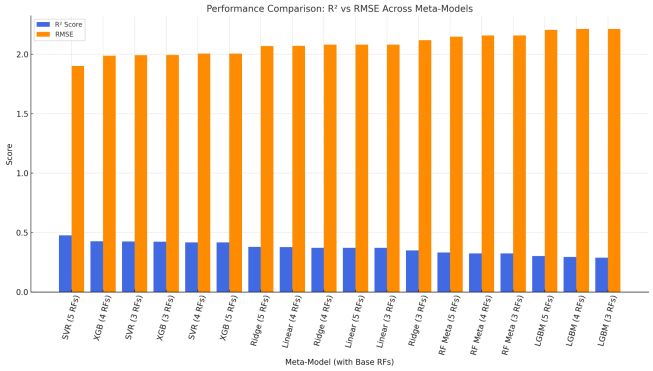
IV. Discussion

The results of this research demonstrates the effectiveness of ensemble learning and improve the prediction of drug delivery efficiency. By implementing stacked RF models, the results reveal the ensemble approaches consistently outperformed individual base models. Both the number of RF base models and the type of meta model impacted the predictive performance

Simple meta models such as linear regression and ridge regression display moderate predictive capabilities specifically when low performing base models were pruned. This highlights the importance of selecting the optimal number of base models in simple algorithms. Nosity predictor can reduce the overall model accuracy. In contrast nonlinear meta-models like Support Vector Regression (SVR) and XGBoost consistently achieved the highest performance across multiple experiments. Specifically the SVR meta-model achieved the highest R^2 score of 0.4842 and the lowest RMSE of 1.8852 when using 10 Random Forest base models.



Linear meta-models such as Ridge Regression exhibited moderate predictive capabilities, particularly when low-performing base models were pruned. This highlights the importance of careful base model selection in linear ensembles, where noisy predictors can reduce overall model accuracy. In contrast, nonlinear meta-models such as Support Vector Regression (SVR) and XGBoost consistently achieved higher performance across multiple experimental configurations. In particular, the SVR meta-model achieved the highest R^2 score of **0.4842** and the lowest RMSE of **1.8852** when using five Random Forest base models.



The results show that when stacked ensembles are paired with proper pruning strategies the models can effectively generalize even on small and high variance datasets. Implementing hyperparameter tuning also proved to be helpful to nonlinear meta-models further enhancing the predictive ability of the models The experimental results also demonstrate that stacked ensembles, when paired with proper hyperparameter tuning and pruning strategies, can generalize effectively even on small and high-variance medical datasets. Hyperparameter tuning proved particularly beneficial for nonlinear meta-models, further boosting predictive stability and reducing model error.

Even though the results of this research demonstrate the effectiveness of ensemble learning when predicting nanoparticle delivery efficiency there are several limitations that need to be considered. Firstly, the size of the dataset is relatively small, only consisting of a few hundred rows of data. The data was incomplete and had lots of new values which restricted the models ability to generalize unseen cases.

For future work expanding the dataset through collecting more data would be highly beneficial. Furthermore exploring more advanced ensemble strategies like deep ensembles could further enhance the predictive accuracy. Finally evaluating the model's performance across different types of biomedical datasets could provide greater insight into its generalizability.

V. Conclusion

Overall this research displays the effectiveness of ensemble learning techniques in improving the prediction of targeted drug delivery. By stacking multiple random forest base models and employing various meta-models I observed significant performance gains over traditional single model approachable. Specifically non-linear meta-models such as SVR and XGboost demonstrated strong performance consistently outperforming more simple models like Linear Regression, achieving the highest R^2 scores and lowest RMSE values across experiments.

This research also highlights the importance of the pruning process. Pruning weaker random forest models led to increased accuracy, specifically when poured with meta modes sensitive to noise. Generating synthetic data also offered an additional method to address limitations of small datasets, further enhancing model accuracy.

Ultimately the best performing stack ensemble consisted of 10 RF base base models and SVR as the meta model. This combination of models achieved a high improvement in the predictions results further suggesting ensemble learning is a powerful tool for improving the prediction of targeted drug delivery

VI. Acknowledgment

Thank you Professor Tomitaka for giving us the opportunity to conduct such interesting research.

VII. References

[1] *"Ensemble Learning Models for Drug–Target Interaction Prediction."* *Briefings in Bioinformatics, Volume 23, Issue 5, September 2022.* DOI: 10.1093/bib/bbac254

[2] **Ma, C., Liu, M., Wang, J., Wang, Y., and He, H .** *"Systematic assessment of nanoparticle delivery to tumors by eight different models."* *Journal of Controlled Release, Volume 374, 2024, Pages 219–229.* DOI: 10.1016/j.jconrel.2023.10.021