CS 4850 – Spring 2025

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

Zachary Mitchell and Nicholas Delarosa

Professor Perry

April 29, 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

Nicholas Delarosal[#2]
College of Computing and Software, Kennesaw State University
1100 South Marietta Pkwy SE, Marietta, Georgia
ndelaros8@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.

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.

The synthetic data was generated using a Generative Adversarial Network (GAN) model implemented through the Synthetic Data Vault (SDV) framework. SDV provides specialized GAN architectures optimized for tabular data, ensuring that the synthetic datasets closely emulate the complex structures and dependencies of real-world datasets. Ensuring the synthetic data was valid in comparison to the original data required multiple statistical and machine learning-based evaluation methods. The Sci-Kit Learn library was utilized extensively due to its robust set of tools for data comparison and analysis. The Kolmogorov-Smirnov (KS) Test was applied to assess the similarity between the distributions of individual numeric columns from the real and synthetic datasets. A smaller KS statistic indicated closer alignment. Additionally, the T-Test was conducted to compare the means of numeric attributes, ensuring that the synthetic data captured central tendencies accurately without introducing significant bias. More statistical tests originating from the SDV library's internal evaluation modules provided additional analysis. The Diagnostic Report analyzed the structural integrity of the synthetic data, focusing on the preservation of inter-column relationships, the replication of missing data patterns, and categorical column accuracy. The Quality Report provided an aggregated quality score that

quantitatively summarized how realistic and useful the synthetic data was when treated as a substitute for real data. These reports evaluated both individual column shapes and pairwise column trends, ensuring that the synthetic dataset preserved not just statistical properties but also relationships between the different data points. Principal Component Analysis (PCA) was used as a dimensionality reduction technique to visually inspect the variance captured by the synthetic dataset compared to the real one. Close clustering in PCA plots indicated that the synthetic data maintained the structural properties of the original dataset across principal components. The hyperparameters selected for tuning the synthetic generation models had profound effects on the resulting data quality. Parameters such as the learning rate, batch size, generator and discriminator hidden layer dimensions, and the number of training epochs were adjusted systematically.

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 we continued to increase/decrease the number of RFs until we 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. 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

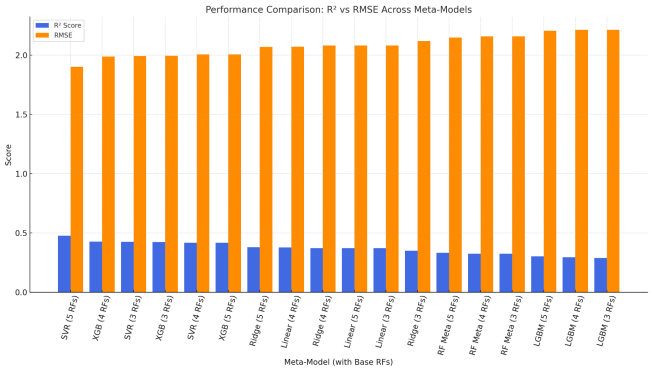
To optimize performance, experiments were conducted changing the number of base models from a range of 3 to 10 RF's and pruning weaker base models when necessary. Hyperparameter tuning was applied to the most efficient model. In the final phase of optimization the most accurate model was retrained using a combination of real and synthetic data to evaluate the impact of synthetic data generation on model performance. The same methodology of base RF models feeding into meta-models was applied to this expanded dataset.

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

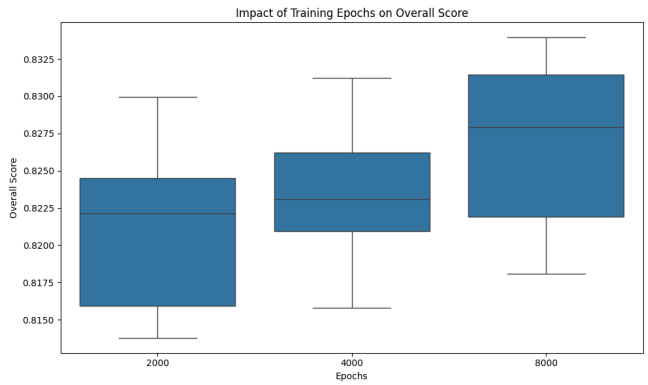
Simple meta models such as linear regression and ridge regression display moderate predictive capabilities specifically when low performing base models were pruned. This highlights the importance of selecting the optimal number of base models in simple algorithms. Nosity predictor can reduce the overall 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.

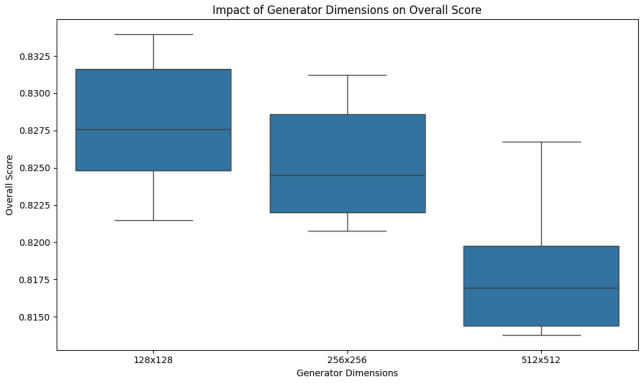


The results show that when stacked ensembles are paired with proper pruning strategies the models can effectively generalize even on small and high variance datasets. Implementing hyperparameter tuning also proved to be helpful to nonlinear meta-models further enhancing the predictive ability of the 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.

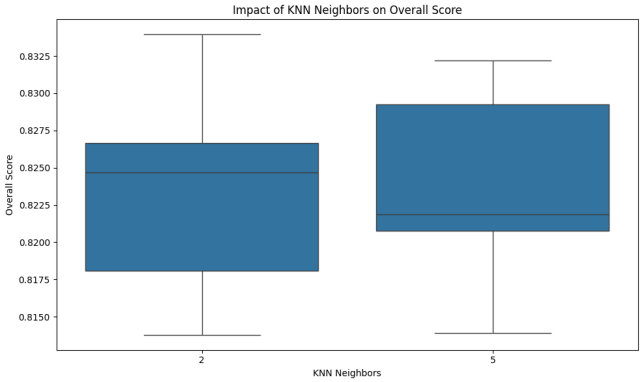
The hyperparameters used for generating the synthetic data helped us generate the best synthetic data we could.The impact of training epochs on the overall synthetic data quality shows that increasing epochs from 2000 to 8000 leads to a slight improvement in the median overall score, with the highest scores observed at 8000 epochs. The trend being increasing the time spent increased the quality of the data.



The generator dimension analysis indicates that smaller generator networks (128×128) consistently produce higher and more stable overall scores compared to larger architectures. As generator size increases to 512×512, the scores noticeably drop and variability increases. This suggests that overly complex generators may overfit, struggle to generalize, and model the tabular data structure inefficiently



The KNN neighbors impact shows a modest difference between using 2 and 5 neighbors during imputation or preprocessing. A lower neighbor count (k=2) slightly outperforms k=5 in median overall score, with tighter distribution, suggesting that less aggressive smoothing or local interpolation better preserves relationships in the data that synthetic models benefit from.



Overall, the results emphasize that moderate complexity and longer training favor better synthetic data quality, while overly large generator networks or excessive smoothing can degrade performance. This means our optimal generation involves higher Epoch, lower generator dimension, and a middling KNN to

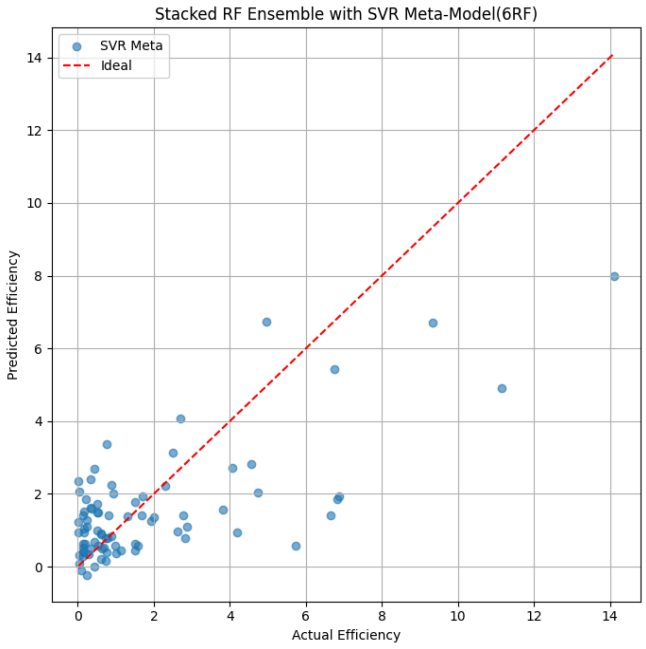
V. Conclusion

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 we 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



While CTGAN is a powerful model for synthetic tabular data generation, other models could also be employed, each offering distinct advantages depending on the project goals. TVAE (Tabular Variational Autoencoder) is another model available in SDV that uses a probabilistic approach, enabling it to handle continuous features more smoothly and often generating better results when the real dataset is highly continuous rather than categorical. Gaussian Copula models provide a fast and interpretable alternative by modeling the distribution of variables through Gaussian transformations, making them effective for smaller datasets or when computational resources are limited. Outside of SDV, Conditional GANs (cGANs) can be adapted to better condition the synthetic data generation on specific attributes, improving control over the output data distribution. These alternatives offers trade-offs: while CTGAN excels at modeling complex feature dependencies and mixed data types, TVAE might outperform it in continuous-heavy datasets, and Gaussian Copula can offer simplicity, speed, and stability when deep learning models like CTGAN are too resource-intensive or prone to overfitting. Selecting the appropriate model ultimately depends on the target application's needs, data structure, and performance evaluation criteria.

.

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