AT1 ML Assisted Nano Bio Tech

CS 4850 - Section 4 – Spring 2025
Dr. Tomitaka
04/28/25

By: Zachary Mitchell

Table of contents

01 Background

What is the Problem?

Traditional Medical Methods

- **Systemic administration:** medication travel through body **without** specific target
- **Lack of precision**: only **small** portion of drug r**eaches** intended site
- Requires high doses: Increase risk of **toxicity** and **side effects**

Targeted Drug Delivery

- **Enhances** delivery of therapeutic agents **directly** to diseased cells
- Uses **nanotechnology**, biomaterials, smart drug carriers
- **Reduces** side effects and **improves** treatment efficacy

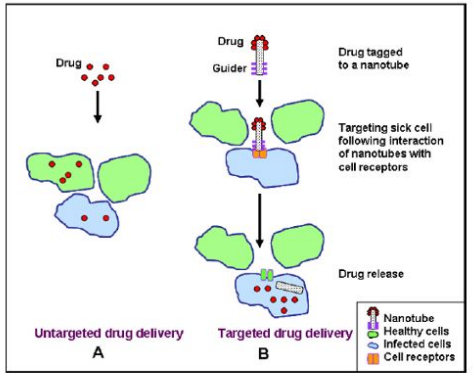


Figure I

1. **Drug Encapsulation**
   - Drugs are encapsulated in nanoparticles, liposomes, or biomaterials.
2. **Targeting & Release**
   - Drugs are directed to diseased cells using biomarkers, ligands, or stimuli like pH, heat, or magnetism
3. **Solution and Role of ML**
   - Solution: ML can analyze datasets to predict optimal drug formulation and improve efficiency

What are we working on?

Main goal:

To improve the prediction of Delivery Efficiency (DE) in nanobiotechnology using machine learning approaches, with the broader aim of advancing data-driven design of nanoparticles for more effective drug delivery.

1. **Ensemble Model**
   - Combine multiple models (e.g., Random Forest, XGBoost, DNN) to create a stacked ensemble

02

Development Methodology

Software Development Process

Research

- Understanding Problem
- Literature review
- Feature Engineering
- Model Implementation

Hypothesis

- Formulate the hypothesis that targeted drug delivery via nanotechnology can improve treatment efficacy

Data-Preprocessing

- Linear Interpolation
- One Hot encoding
- Z score
- MaxAbsScaler

Model Experimentation

Implement various ML models:

- RandomForest
- XGB
-

Evaluation/ Validation

Evaluate models using metrics such as:

- R Squared
- Mean Squared Error

03

Requirements and Tools

Requirements and Tools

- Machine Learning Models: Ensemble Model, Deep Neural Network,
- Nanoparticle dataset from literature search
- Preprocessing: Need to do the same preprocessing steps
    - Amending missing values & incomplete entries
- Tools & Libraries: Google Colab Python, Scikit-learn, SDV

No.	Type	MAT	TS	CT	TM	Shape	Size
1	INM	Gold	Active	Cervix	Xenograft Heterotopic	Rod	NA
2	INM	Gold	Passive	Colon	Allograft Heterotopic	Rod	NA
3	INM	Gold	Active	Brain	Xenograft Heterotopic	Spherical	1.6232493
4	INM	Gold	Passive	Brain	Xenograft Heterotopic	Spherical	1.5831988
5	INM	Gold	Active	Breast	Xenograft Heterotopic	Others	1.8350561
6	INM	Gold	Active	Skin	Xenograft Orthotopic	Spherical	1.6946052
7	INM	Gold	Active	Skin	Xenograft Orthotopic	Spherical	1.7781513
8	INM	Gold	Active	Skin	Xenograft Orthotopic	Spherical	2.0017337
9	INM	Gold	Passive	Skin	Xenograft Orthotopic	Spherical	1.665581
10	INM	Gold	Passive	Skin	Xenograft Orthotopic	Spherical	1.807535
11	INM	Gold	Passive	Skin	Xenograft Orthotopic	Spherical	2.0178677
12	INM	Gold	Active	Skin	Xenograft Orthotopic	Spherical	2.2445245
13	INM	Gold	Passive	Skin	Xenograft Orthotopic	Spherical	2.2201081
14	INM	Gold	Passive	Brain	Xenograft Heterotopic	Spherical	0.7831887
15	INM	Gold	Passive	Prostate	Xenograft Heterotopic	Spherical	NA

DE_tumor	DE_heart	DE_liver	DE_spleen	DE_lung	DE_kidney
1.971937133	0.269387786	26.41061294	2.71416881	0.294177128	0.854130379
1.278030303	NA	NA	NA	NA	NA
2.821588295	0.317316655	11.54321037	4.383887921	0.420119952	1.16313958
3.186238491	NA	NA	NA	NA	NA
0.733884298	0.224474552	41.42407658	1.93159841	0.765849448	1.935310941
24.93100559	NA	NA	NA	NA	NA
23.53137528	NA	NA	NA	NA	NA
20.88493965	NA	NA	NA	NA	NA
18.57177319	NA	NA	NA	NA	NA
14.14878998	NA	NA	NA	NA	NA
11.15403219	NA	NA	NA	NA	NA
8.085157857	NA	NA	NA	NA	NA
4.927649187	NA	NA	NA	NA	NA
1.128149606	NA	NA	NA	NA	NA

04 Research

Ensemble Models

Main goal:

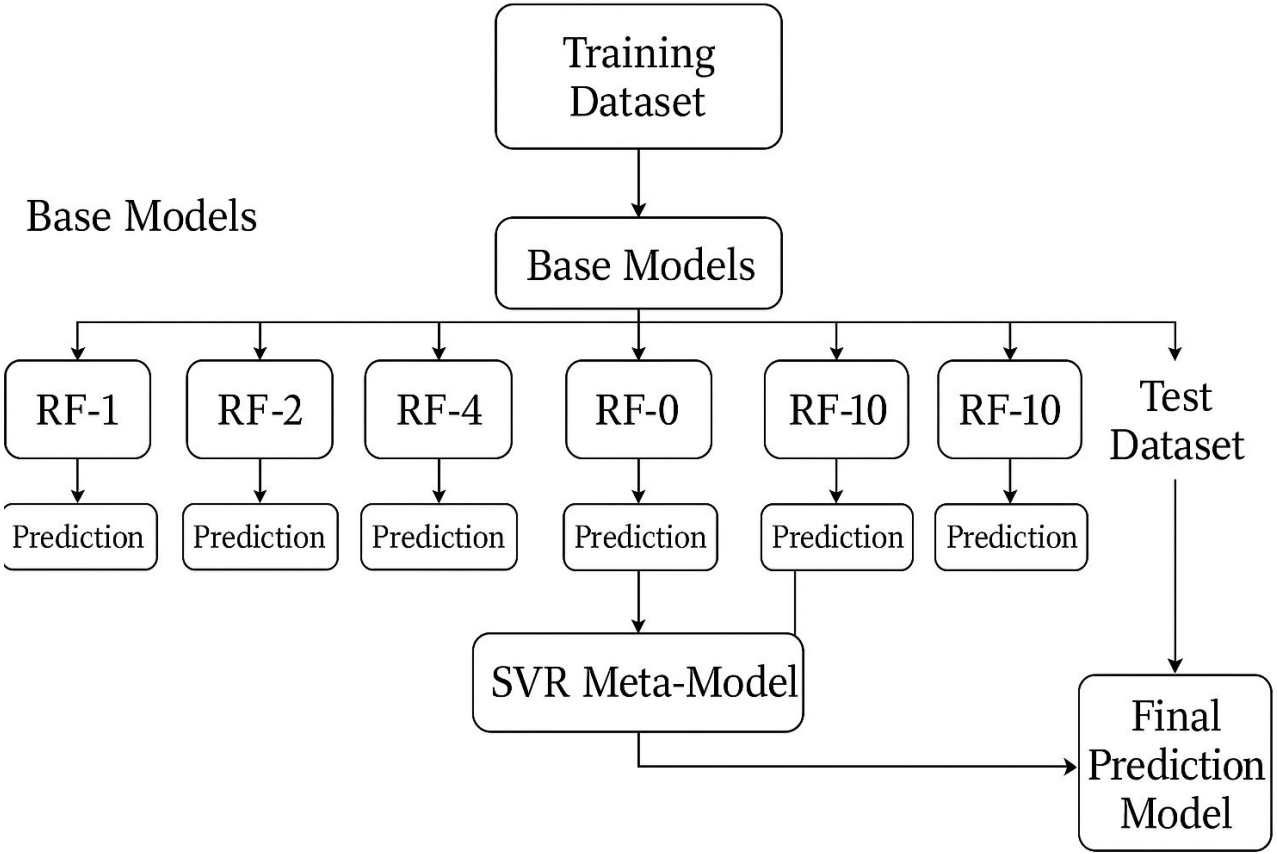
- Evaluate the effectiveness of 5 different meta-models (SVR, XGB, Ridge, Linear, RF, LGBM) stacked on top of 3–5 Random Forest base models for predicting targeted drug delivery efficiency
- Investigate how synthetic data affects performance of ensemble models

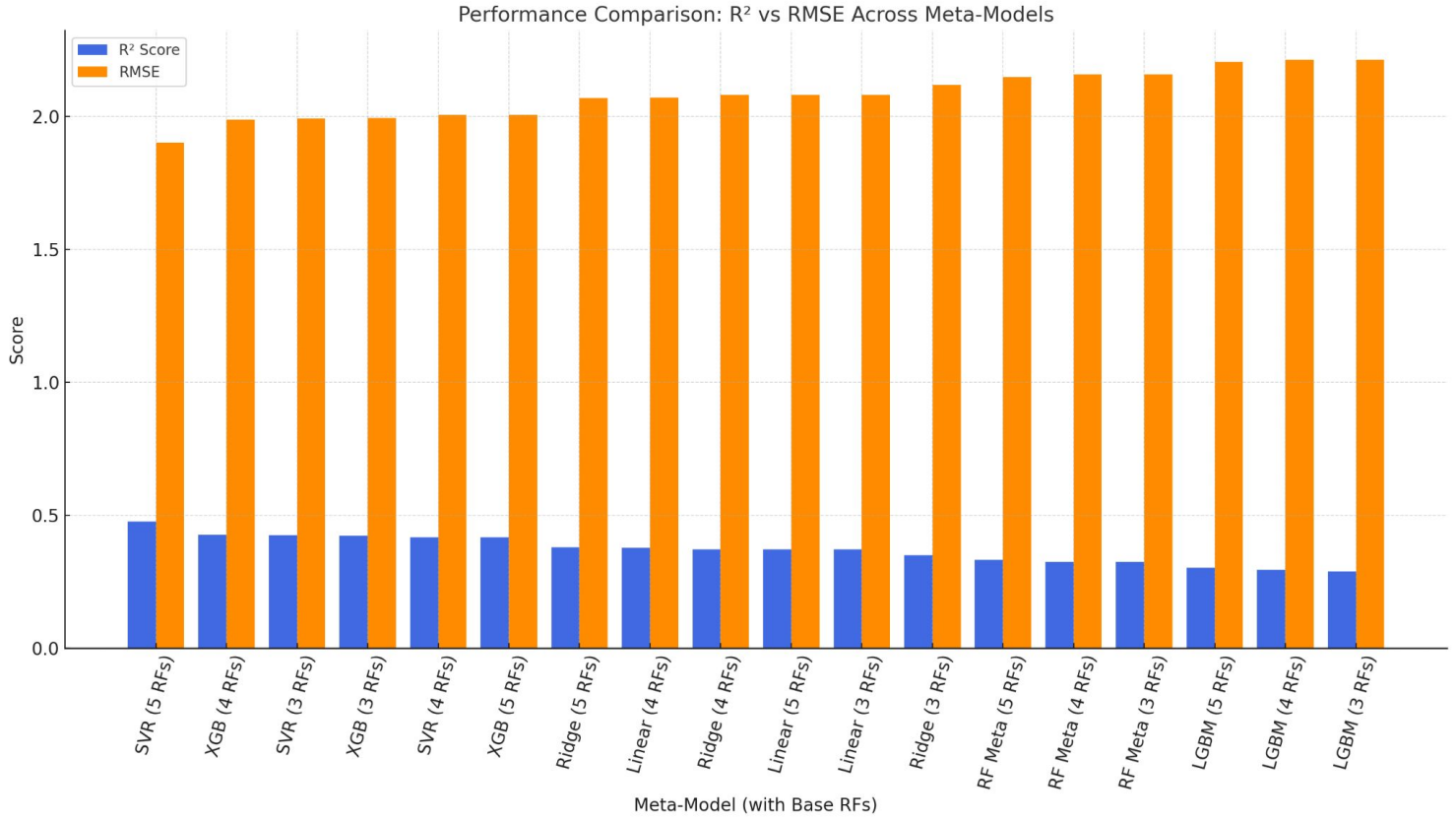
Process Overview

1. Train multiple Random Forest base models on the training data (real + synthetic)
2. Generate predictions from each base model
3. Use base model predictions as input to train the meta-model
4. Evaluate meta-model performance on test data using R^2 Score and RMSE.

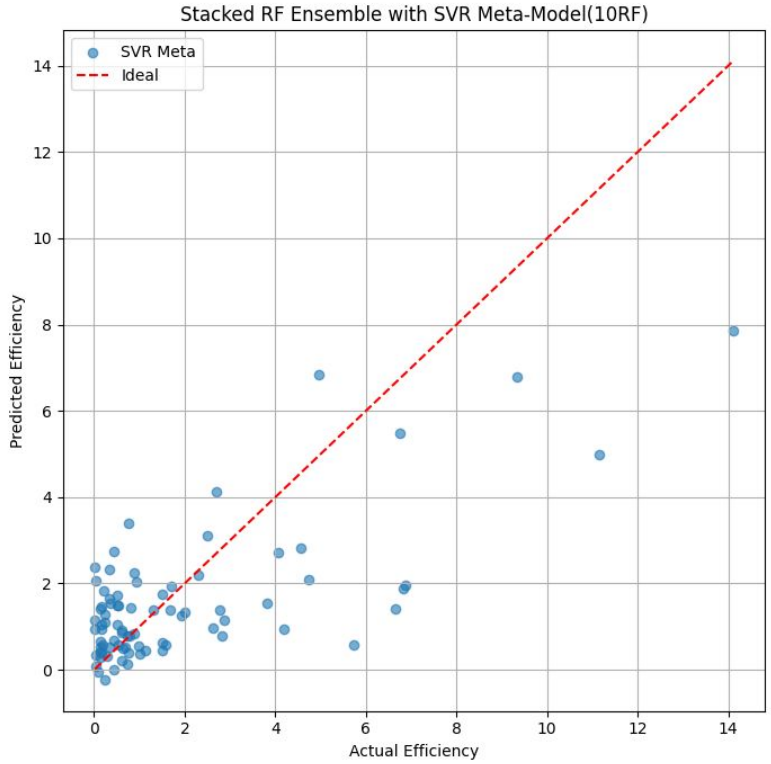
Extra Information

- Bagging (bootstrap aggregation) was used to create diverse RF base models
- Both linear and nonlinear meta-models were evaluated.


Training Dataset
Base Models
Base Models
RF-1
RF-2
RF-4
RF-0
RF-10
RF-10
Test Dataset
Prediction
Prediction
Prediction
Prediction
Prediction
Prediction
SVR Meta-Model
Final Prediction Model


Performance Comparison: R² vs RMSE Across Meta-Models
R² Score
RMSE
Score
0.0
0.5
1.0
1.5
2.0
SVR (5 RFs)
XGB (4 RFs)
SVR (3 RFs)
XGB (3 RFs)
SVR (4 RFs)
XGB (5 RFs)
Ridge (5 RFs)
Linear (4 RFs)
Ridge (4 RFs)
Linear (5 RFs)
Linear (3 RFs)
Ridge (3 RFs)
RF Meta (5 RFs)
RF Meta (4 RFs)
RF Meta (3 RFs)
LGBM (5 RFs)
LGBM (4 RFs)
LGBM (3 RFs)
Meta-Model (with Base RFs)

Top Performing Model



R2: 0.4842

RSME:1.8852

Results Analysis

Key takeaways:

- SVR (5 RFs) outperformed all meta-models with an R2 Score of 0.4755 and the lowest RMSE of 1.9012, indicating it best captured the patterns in the data while minimizing error.
- XGB with 4 RFs closely followed with an R2 of 0.4258 and RMSE of 1.9891.
- Linear and Ridge meta-models, while simpler, performed decently but fell short in predictive power compared to nonlinear models like SVR and XGB.

Interpretation:

- Non-linear models (SVR, XGB) generalized better, especially when all 5 RF base models were included.
- Using more RF base models generally improved stability and prediction quality, but there are diminishing returns after 4 models
- Overfitting was controlled well, as seen in similar RMSE across models with different numbers of base learners.
-

05

Limitations + Future Applications

Requirements and Limitations

Projects Limitations:

- Data Size: Limited number of samples may reduce model generalizability
- Imbalanced Dataset: Some outcomes and materials are underrepresented
- Overfitting: Risk due to high model complexity especially with DNS
-

Future Work:

- Implement synthetic generated dataset into and use to make models
- Extend framework to predict delivery efficiency for other diseases or delivery routes
- Develop web application that takes nano properties as inputs and predicts delivery efficiency using different models

06

Conclusion

Conclusion

- Research explores how Machine Learning models can improve the prediction of drug delivery efficiency for targeted nanomedicine
- Through experimentation and model comparison we found that so single model is perfect but combining them increases robustness and prediction power

Any Questions?