“End-to-End ML pipeline for Warfarin Dosing Prediction”

Group Members:

Zaheer Khan

Tanisha Londhe

Major: Bioinformatics and Computational Biology

Course: Introduction to Machine Learning (CSCI-4750-01)

Instructor: Dr. Jie Hou

Saint Louis University

Abstract:

The present study examines the machine learning models to forecast a perfect right warfarin therapeutic dose for patient data, having a pathological or genetic background. Warfarin dosing is highly variable and is patient specific (varies according to age, weight, genotype, and desired INR). We deployed a machine learning pipeline which involves data preprocessing, multicollinearity analysis, model training and corresponding hyperparameter tuning, and model performance. For the tested models Random Forest and XGBoost produced the best prediction of warfarin dose. This project illustrates the way ML can promote the improvement of clinical decision-making and drug personalization.

Introduction:

Warfarin is an orally administered anticoagulant widely used in the prevention and treatment of thromboembolic events in atrial fibrillation, deep vein thrombosis, and management of the prosthetic heart valve. Spite being effective, Warfarin carries a major dosing problem with its narrow therapeutic index and wide variations in dose requirements among people. The (clinical) characteristics of the patient, such as age and weight, concomitant medications, lifestyle factors, and most importantly genetic polymorphisms, such as CYP2C9 and VKORC1, play a role in this variability. Inappropriate dosing may cause serious adverse effects, i.e., excessive bleeding or thromboembolic complications.

Traditionally, clinicians use trial and error methods to get the best dosage of Warfarin, this means monitoring of dosage levels and adjustments are necessary often. This is resource-intensive, time and patient safety risks prone. Therefore, there has been an increasing trend of research in pharmacogenomics and precision medicine in this field to develop predictive models used to estimate optimized Warfarin dosing information for the individual patient.

The present paper summarizes a few research papers, such as International Warfarin Pharmacogenetics Consortium (IWPC), that have studied statistical and machine learning (ML) models to develop Warfarin dose prediction model from demographic, clinical and genetic factors. Computer models of learning provide opportunities to identify nonlinear relationships in data that might be used to enhance prediction power over traditional linear regression models.

For this project, we used the IWPC dataset for developing and comparing several ML models such as Linear Regression, Lasso, Random Forest, Support Vector Regression, XGBoost, and an Artificial Neural Network (ANN) as models for estifying therapeutic Warfarin dose. The project pipeline consists of data preprocessing, exploration and modeling (training and evaluation), feature importance analysis, and deployment of the best models using Gradio web-interface. This report describes the methodology, results and analysis, as well as deployment approach for the Warfarin Dose Prediction task to demonstrate the way machine learning can improve precision dosing in medical practice.

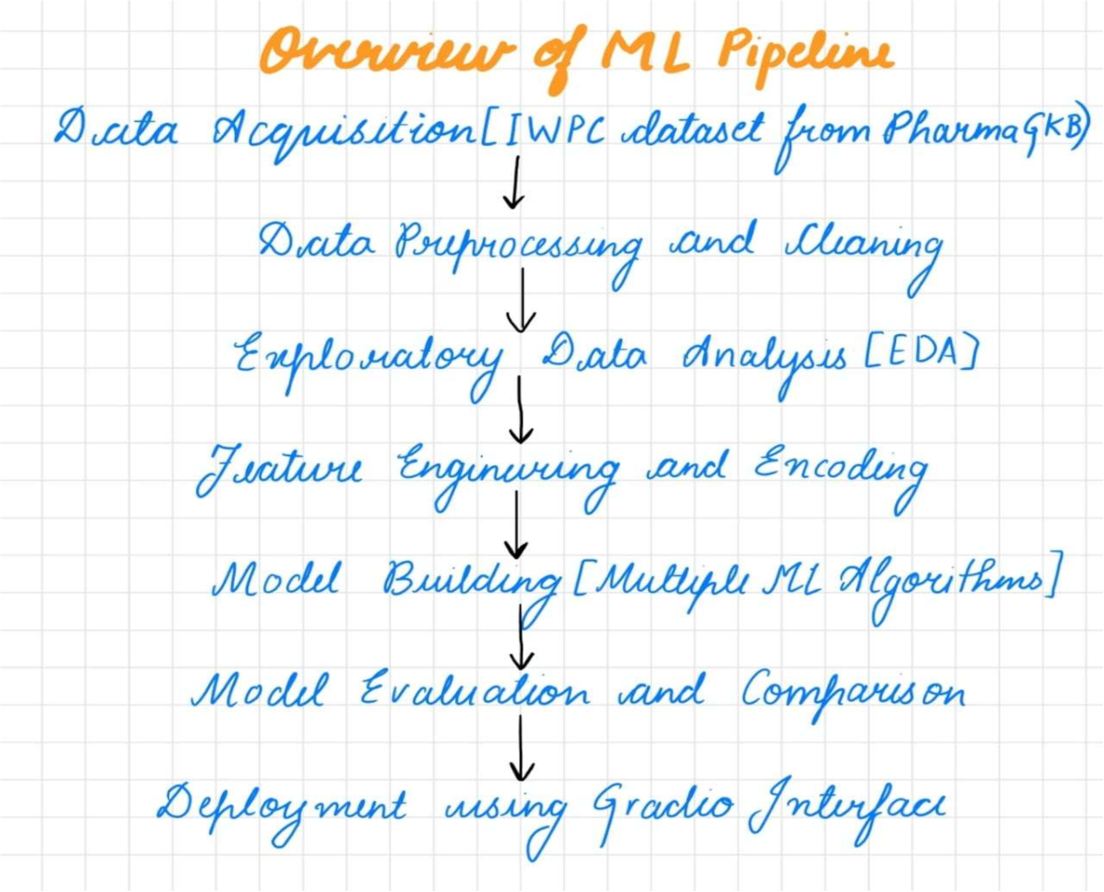
Related Work:

Prior studies by the International Warfarin Pharmacogenetics Consortium (IWPC) have shown that combining genetic as well as clinical information improves the precision of warfarin dose predictions by a large margin. Follow-up research has benefited from a variety of regression and ensemble learning methods to observe continued improvement in the MAE and R^2 key performance measurements. Even though deep learning is not fairly adopted yet in this area, it promises a rich nonlinear model of predictors. In this project, we extend these prior fundamental efforts by introducing the use of many traditional algorithms, an artificial neural network (ANN), strict feature selection procedures, and a thorough evaluation paradigm.

- IWPC (2009): designed a regression–based dosing algorithm, incorporating clinical as well as genetic factors (NEJM, 360:753–64).
- Hindawi et al. (2015): recast the prediction task as binary classification problem (dose>30 mg/week vs. <=30 mg/week) and evaluated models including logistic regression, support vector machines (SVM), random forests leading to the AUC-ROC of almost 0.85.
- PLOS ONE (2018): the study has used gradient boosting ensemble models and observed better RMSE values for nonlinear models against the standard linear ones.
- Recent deep learning studies (2020): are starting to look for neural network approaches for warfarin dose prediction while difficulties still exist in interpreting model decision because of the lack of model features clarity.

Methodology:

Overview of your experimental pipeline:


Overview of ML Pipeline
Data Acquisition[IWPC dataset from PharmaGKB)
Data Preprocessing and Cleaning
Exploratory Data Analysis [EDA]
Feature Engineering and Encoding
Model Building [Multiple ML Algorithms]
Model Evaluation and Comparison
Deployment using Gradio Interface

Data Acquisition and Preprocessing

The International Warfarin Pharmacogenetics Consortium (IWPC) provided the dataset which PharmGKB database made accessible. The dataset contains clinical information from more than 5000 patients who have variables such as gender, race, age group, height, weight, comorbidities, concurrent medications and genetic markers (CYP2C9, VKORC1). The target variable ‘Therapeutic Dose of Warfarin’ (in mg/week) represents the actual stable weekly dose given to each patient. The ‘Therapeutic Dose of Warfarin’ column was used as the output variable (target) and was not included in the input features used for model training. The first step in preprocessing involved renaming some columns for better understanding and uniformity (e.g., ‘Therapeutic Dose of Warfarin’ was renamed to ‘Dose’). The data cleaning process used median imputation for numerical features (such as Height, Weight, INR Therapeutic) and mode imputation for categorical features (Gender, Race, Diabetes, Simvastatin, Amiodarone, Target INR) to handle missing values. These steps resulted in a complete dataset that was ready for analysis and modeling.

Feature Engineering and Encoding

The categorical features Gender, Race, Age group, CYP2C9 genotype, VKORC1 genotype received One-Hot Encoding to transform them into binary dummy variables. StandardScaler transformed the numerical features Height, Weight, Target INR, and INR Therapeutic to achieve uniform feature scaling because SVR and ANN models depend on feature magnitude. The Dose values underwent log-transformation using np.log1p to address the target variable distribution skew which improved regression performance.

Exploratory Data Analysis (EDA)

Some in-depth explorations of dataset need to be performed in order to get insights into the dataset and see how variables are related to one another – extensive Exploratory Data Analysis (EDA) was performed. The investigation of distribution of the target variable, Therapeutic Dose of Warfarin (mg/week), was carried out first. The distribution was right-skewed as presented by the histogram below; most patients would therefore require low doses each week while a few should receive much higher doses weekly. Such skewness can affect negatively performance of regression models through its violation of normality. To minimize this, a logarithmic transformation came in to normalize the target distribution in viewing for modeling.

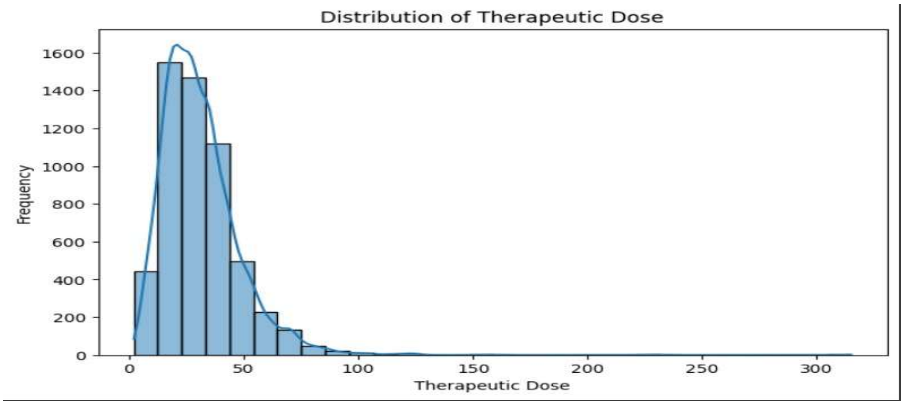


Figure 1: Distribution of Therapeutic Dose

A correlation heatmap of numerical features was generated to examine inter-feature relationships and assess potential multicollinearity. A strong positive correlation was observed between Height and Weight (correlation coefficient = 0.52), as expected due to their biological linkage. Importantly, INR Therapeutic showed a moderate correlation with Warfarin Dose (correlation coefficient = 0.18), reaffirming its significance in dose prediction. Overall, no extreme multicollinearity was detected, allowing inclusion of all features in model development.

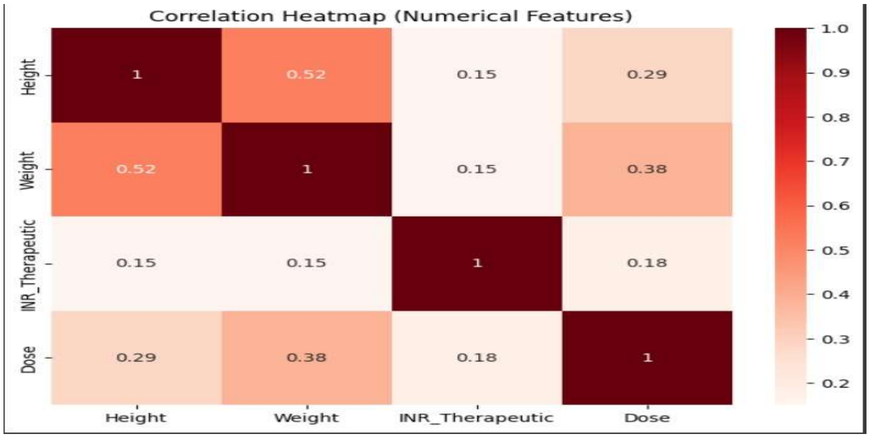


Figure 2: Correlation Heatmap

Further, subgroup visualizations were applied to check how Warfarin Dose varied across Age groups, Gender and Race categories. Boxplot of Dose by Age group exhibited a generalized tendency in which middle-aged patients (40-49, e.g.) experienced larger doses requirements while older groups (80-89, 90+) showed lower requirement. This finding is consistent with clinical information showing an age-related increase in Warfarin sensitivity.

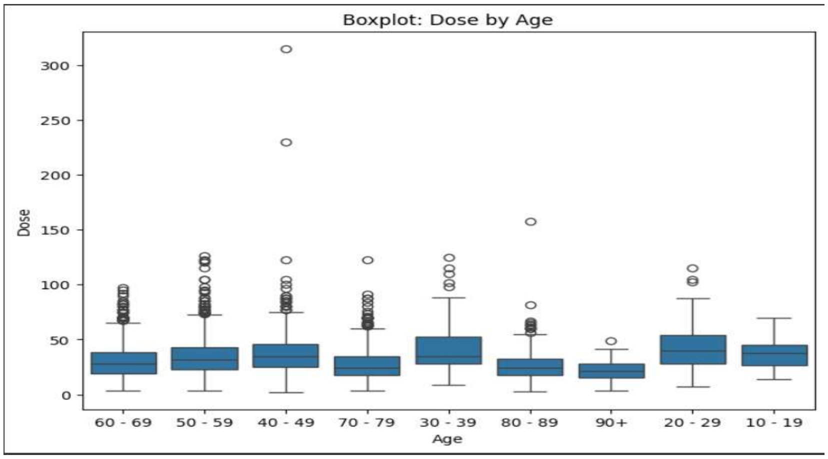


Figure 3: Boxplot of Dose by Age group

Gender-wise dose distribution was examined using a violin plot, illustrating a marginally higher median dose for male patients compared to females. However, the overall distribution shapes were quite similar, suggesting gender has a moderate influence on Warfarin dose variability.

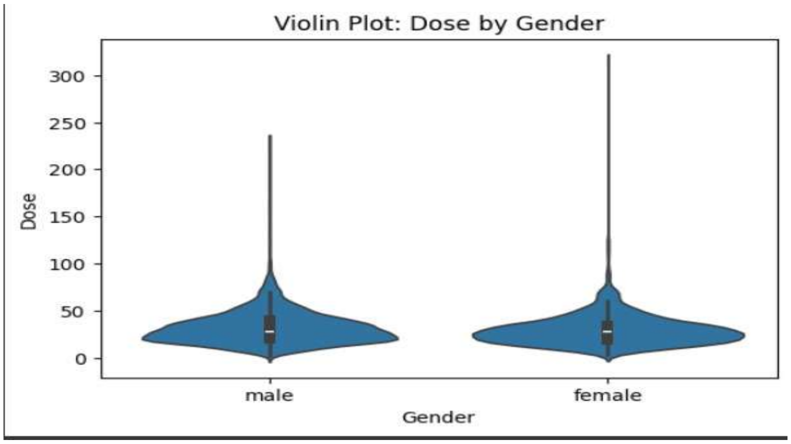


Figure 4: Violin Plot of Dose by Gender

Finally, a strip plot of Dose by Race was used to examine between-racial variations. The plot emphasized the fact that particular racial groups, such as White and Caucasian had a broader distribution of dose values; whereas other, like Japanese and Han Chinese grouped at lower dose levels. Such differences underscore the importance of the inclusion of race-specific genetic and clinical factors into predictive modeling.

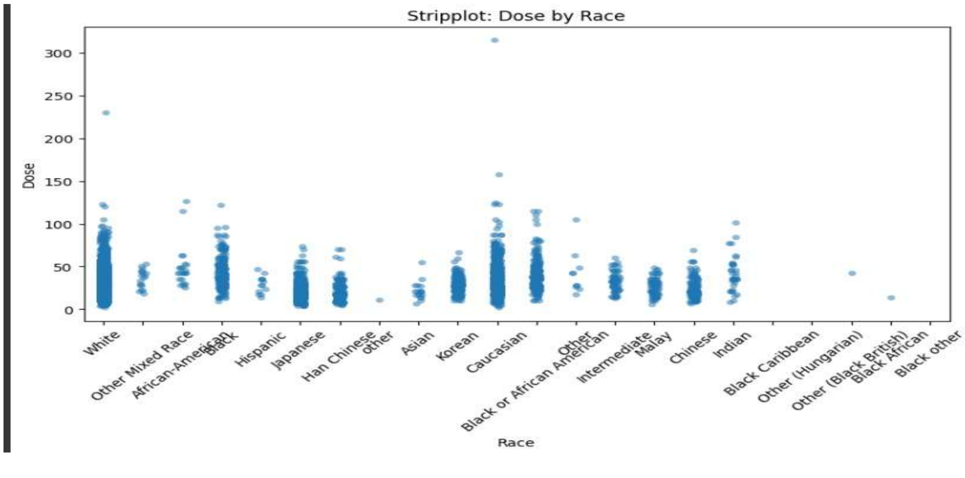


Figure 5: Strip Plot of Dose by Race

These exploratory analyses yielded useful clinical and statistical insights that took the next feature selection and model building steps in this direction. As patterns in Age, Gender, and Race groups were justified upon inclusion as a predictive measure for Warfarin dosing.

Multicollinearity Assessment

Multicollinearity was examined through the correlation heatmap. Although certain features exhibited correlation (e.g., Height-Weight), none exceeded critical thresholds that would necessitate removal. Additionally, models like Lasso Regression were used to perform regularization-based feature selection, while tree-based models like Random Forest and XGBoost inherently managed multicollinearity during training.

Input Features and Target Variable

The final set of input features (X) included Gender, Race, and Age group (one-hot encoded), Height and Weight (scaled), and binary clinical variables such as Diabetes, Simvastatin, and Amiodarone. Coagulation factors like Target INR and INR Therapeutic were included as numerical features. Genetic markers CYP2C9 and VKORC1 were also one-hot encoded to account for genotype variations. The output variable (y) was the Therapeutic Dose of Warfarin (mg/week), used exclusively as the prediction target after applying a log transformation to normalize its distribution.

Machine Learning Models and Algorithms

Multiple regression algorithms were implemented to model the dose-response relationship:

- Linear Regression served as the baseline model.
- Lasso Regression (L1 penalty) was applied for embedded feature selection and shrinkage.
- Support Vector Regression (SVR) with RBF kernel aimed to model complex non-linear patterns.
- Random Forest Regressor utilized bootstrapped decision trees for robust predictions.
- XGBoost Regressor offered optimized gradient boosting with regularization.

- Artificial Neural Network (ANN) with two hidden layers (64 and 32 nodes) and dropout regularization was implemented as the deep learning approach.

Results:

Model Evaluation:

To asses and contrast predictive performance of different regression models several metrics were calculated on the validation and test datasets. These were; Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R² (coefficient of determination). To generate those predictions, the models were constructed first on log- transformed dose values and then Back -transformed in the original dose scale using exponential functions.

For all models, the best performance was Linear Regression, with a Test R² score of 0.43, which suggests it could explain 43% of the variance in Warfarin dose between patients. Its actual vs predicted scatter plot revealed a reasonably tight bunch of points near the red reference line — the ideal place for predictions. The spreading improves with increasing measurements of the dose, but the general trend describes the line well, i.e. accurately.

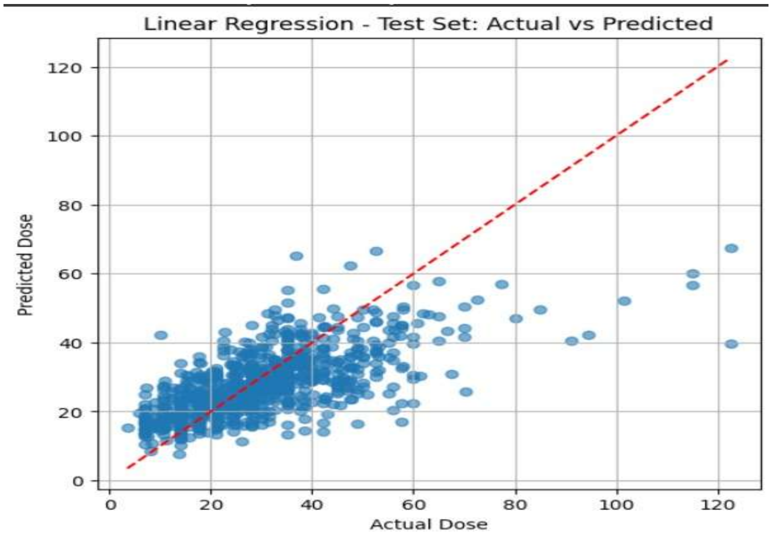


Figure 6: Linear Regression Actual vs Predicted scatterplots

The lowest R² value was obtained by Lasso Regression that uses L1 regularization. Its scatterplot showed a strong inclination to predict middle level doses, as most of the points were bunched horizontally near 30 mg/week. This underperformance is probably because of over penalization and coefficients dampening of features curtailing the ability of the model to discriminate patients in a patient-specific fashion.

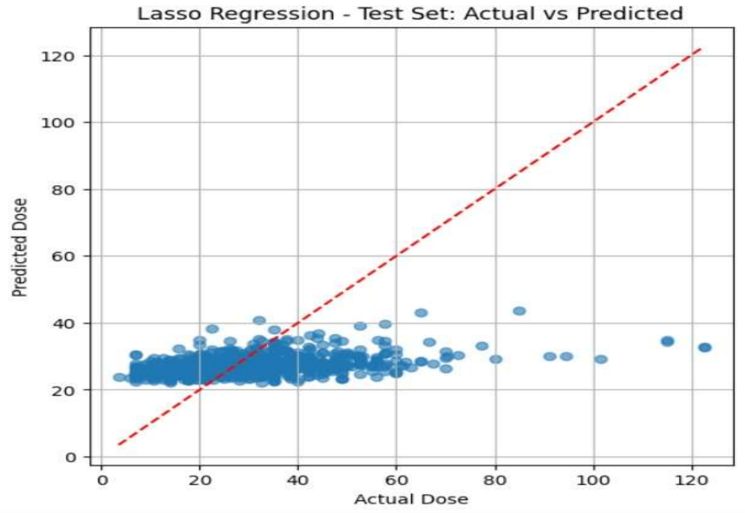


Figure 7: Lasso Regression Actual vs Predicted scatterplot

Random Forest Regressor did not perform very well but averagely well as it explored non – linear patterns using its collection of decision trees. Its plot demonstrated predictions line-up more with the ideal line, but with greater spread, particularly for patients who took larger actual doses. Inspite of that, its R² score for calculating the NOISE was remarkably better than Lasso and SVR, indicating its stability to the noises and interactions of variables.

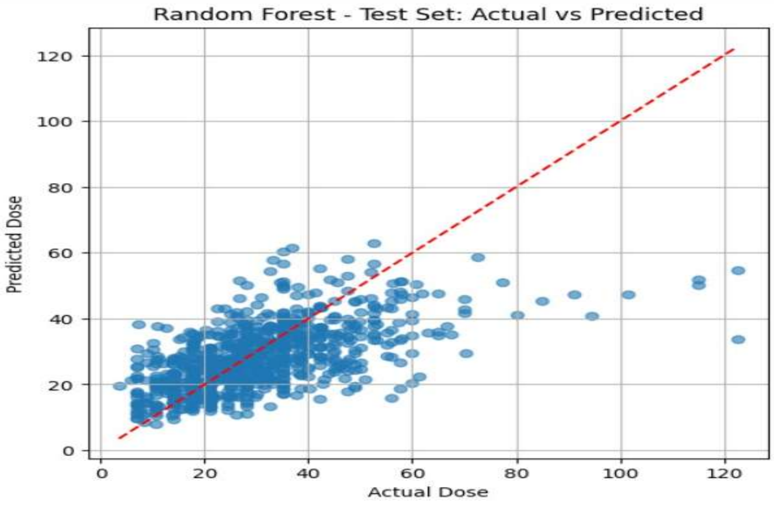


Figure 8: Random Forest Actual vs Predicted scatterplot

On the other hand, the Support Vector Regression (SVR) model was severely underperforming, in terms of prediction accuracy and visual alignment. The SVR plot spread was significant and non homogeneous from the diagonal indicating that the model had difficulty in generalizing over the full range of values of dose. Its tendency to under predict high-dose patients and over predict low-dose ones is most likely what lead to its lower performance metrics.

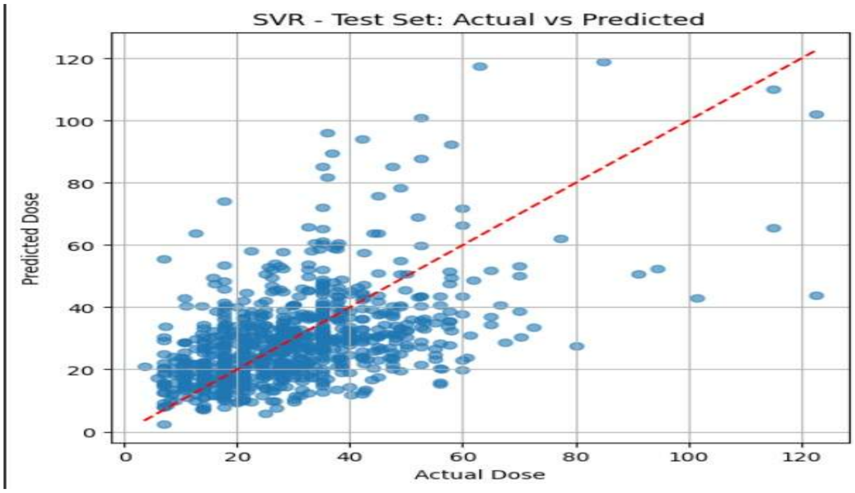


Figure 9: SVR Actual vs Predicted scatterplot

Finally, the XGBoost Regressor demonstrated excellent performance only second to Linear Regression, with a close to exact fits for the actual doses with little error. The scatter diagram of predictions coordinates well with the best reference line, with few outliers, particularly for middle-range doses. In this performance, it demonstrates XGBoost's strength in terms of the skill in capturing complex feature interactions and reducing the bias-variance tradeoff.

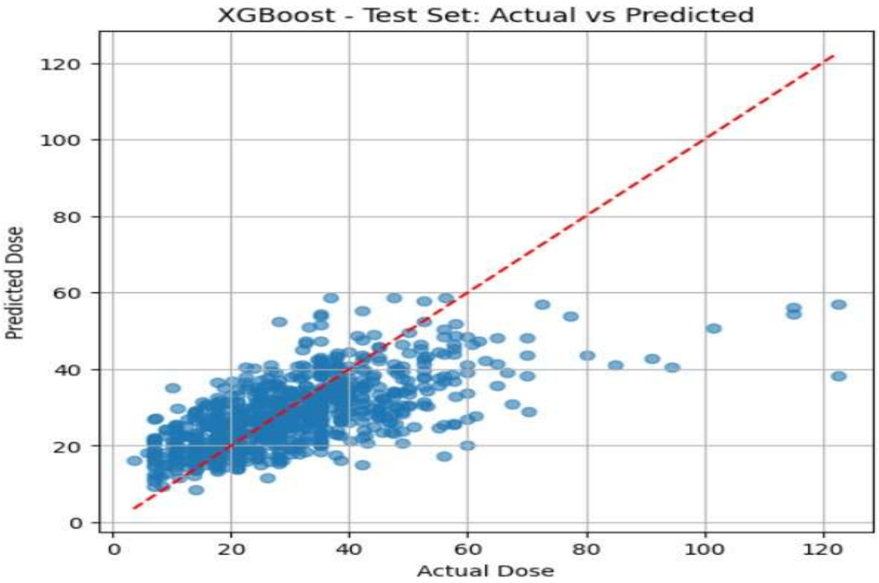


Figure 10: XGBoost Actual vs Predicted scatterplot

Overall, Linear Regression and XGBoost were the most effective models, both in terms of statistical metrics and visual accuracy. Random Forest provided moderately good predictions, while Lasso and SVR models displayed limited predictive capacity. These evaluations informed the selection of models to be included in the final Gradio-based deployment interface.

Deep Learning Model:

To explore the potential of deep learning in Warfarin dose prediction, an Artificial Neural Network (ANN) was developed and trained on the processed dataset. The ANN architecture consisted of two hidden layers with ReLU activation functions and dropout regularization to prevent overfitting. The model was trained using the Adam optimizer with mean squared error as the loss function.

The training process was monitored through the ANN Loss Curve, which plotted both training and validation losses over 50 epochs. As depicted in the figure, both loss curves exhibited a steep decline during the initial epochs, stabilizing as the model converged. Importantly, the validation loss closely followed the training loss without significant divergence, indicating effective generalization and absence of overfitting.

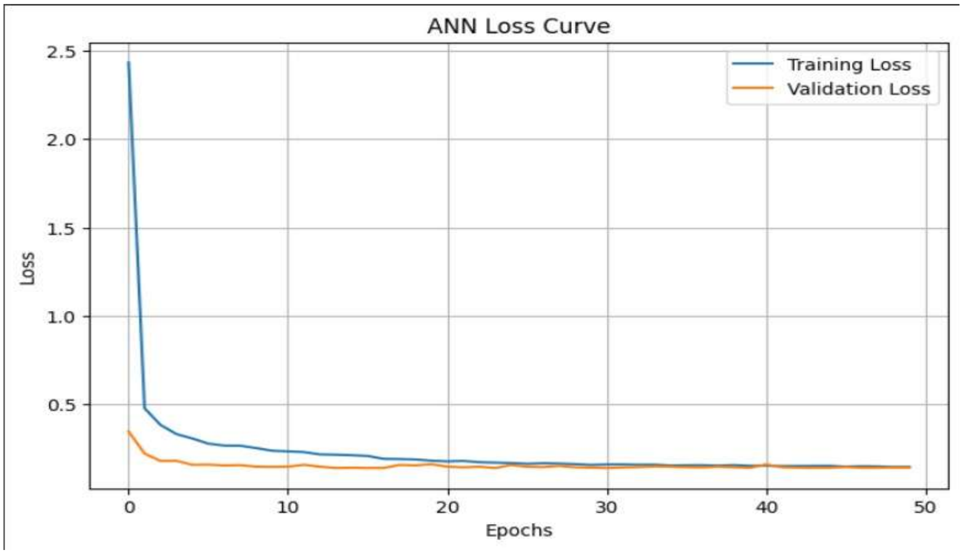


Figure 11: ANN Loss Curve

The ANN's predictive performance on the test set was visualized through an Actual vs Predicted scatterplot. The plot demonstrated that the ANN was able to approximate the target dose values reasonably well, with most predictions aligning along the diagonal reference line. While the spread was slightly wider compared to linear models, the ANN captured non-linear relationships better for certain patient subsets, especially in the mid-dose range.

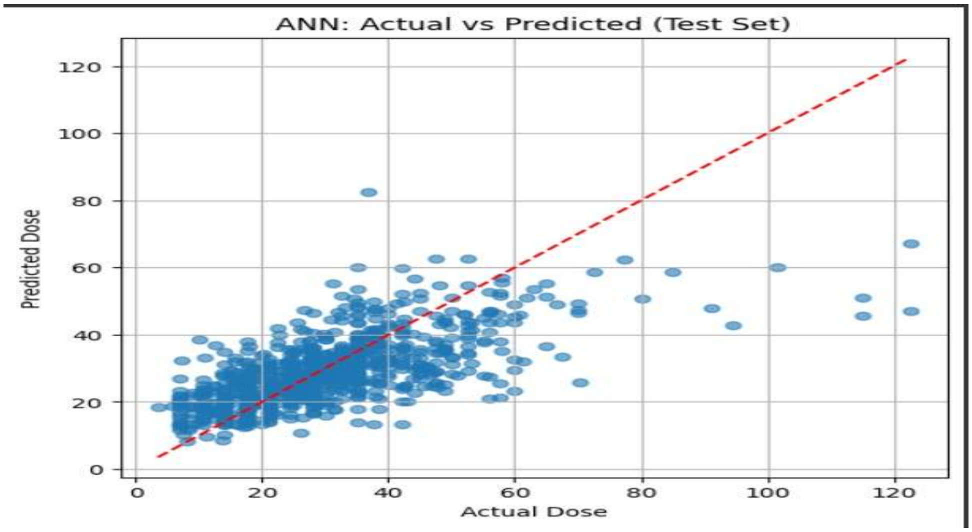


Figure 12: ANN Actual vs Predicted scatterplot

Overall, the ANN model delivered competitive performance, comparable to tree-based methods like Random Forest and XGBoost. Its ability to model complex, non-linear interactions highlights the relevance of deep learning approaches in precision dosing applications, though simpler models like Linear Regression still outperformed in interpretability and marginally in accuracy.

Model	Val MAE	Val RMSE	Val R²	Test MAE	Test RMSE	Test R²
Linear	9.05	15.58	0.31	8.22	11.67	0.43
Lasso	11.38	18.24	0.05	10.70	14.76	0.09
SVR	12.26	19.38	-0.07	11.09	15.16	0.04
Random Forest	9.78	16.36	0.23	8.83	12.50	0.35
XGBoost	9.13	15.89	0.28	8.31	11.81	0.42
ANN	9.15	13.02	0.31	8.50	12.10	0.38

Table 1: **Final Model Results**

Feature Selection and Dimensionality Reduction:

It is important for clinical interpretability to know which of the features contribute most to Warfarin dose prediction. The feature relevance was estimated using two complementary methods. Random Forest Feature Importances Lasso Regression Coefficients.

The most important predictor in Random Forest Feature Importance plot was Weight, followed by INR Therapeutic and Height. Other genetic predictors such as VKORC1_G/G genotype and CYP2C9 also arose as great contributors. This is consistent with pharmacogenetic evidence, whereby, these polymorphisms are associated with Warfarin metabolism and sensitivity. In addition, demographic variables such as Age group (70-79), Gender had moderate importance indicating their clinical implications in dosing decisions.

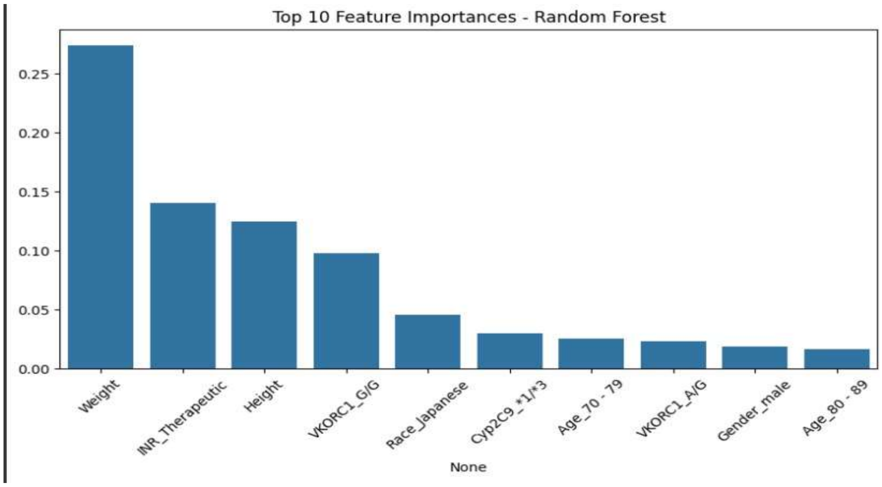


Figure 13: Random Forest Feature Importance plot

While Lasso Regression, which conducts embedded feature selection via L1 regularization chose only Weight and Height as important predictors with nonzero coefficients. The absolute value of Weight's coefficient dwarfed that of Height confirming its leading role in establishing dose variability. Such a sparsity in the selection of Lasso implies its harsh regularization, which may overlook weaker but clinically important features.

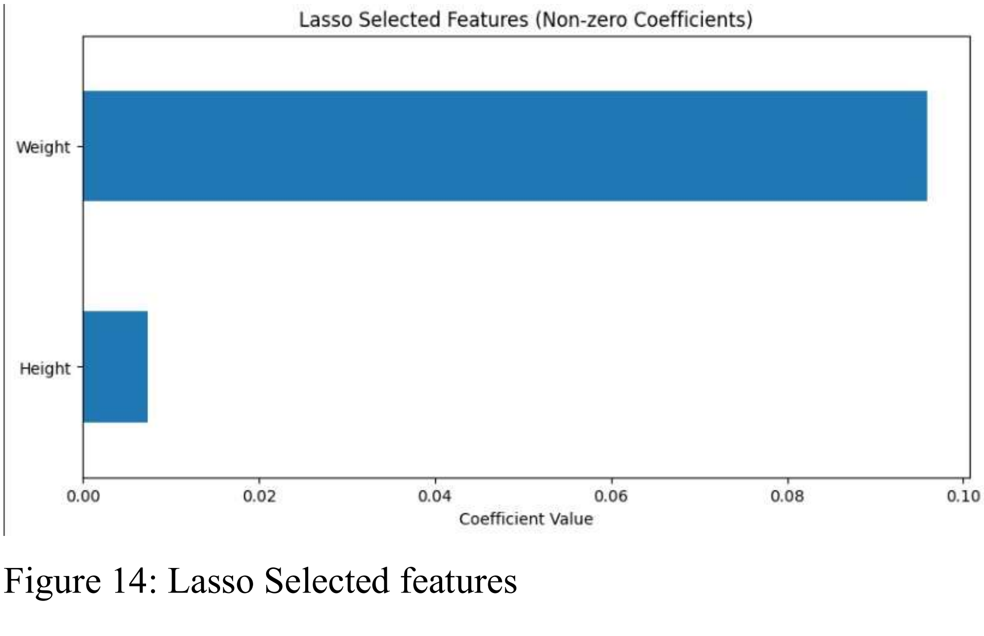


Figure 14: Lasso Selected features

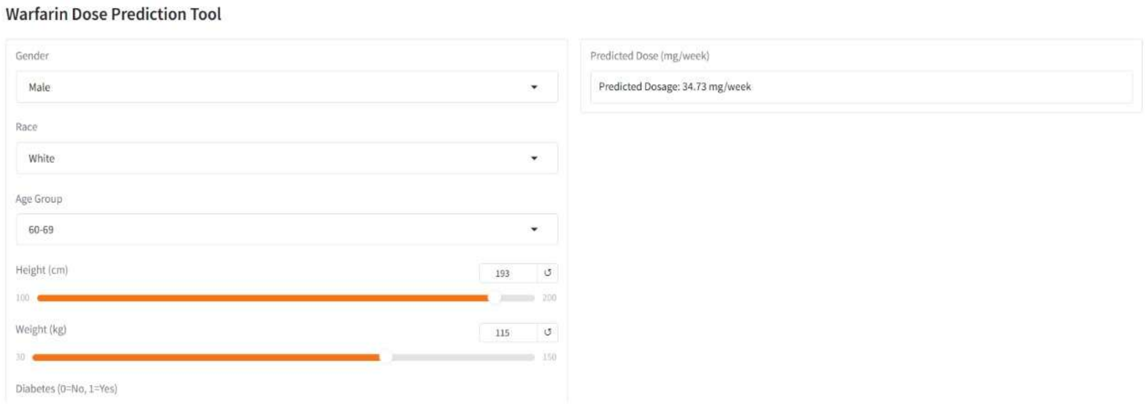
By integration of clues from Random Forest and Lasso analyses, Clinical variables (Weight, INR, Height) and Genetic factors (VKORC1, CYP2C9) are prominently responsible for requirements of Warfarin dosage requirements. Although Random Forest describes complex feature interactions, Lasso provides a compromise between the accuracy and interpretability of the model due to paying attention to the strongest predictors.

Model Deployment via Gradio Interface

To turn the machine learning models developed into a practical clinical tool, an internet based gradio interface was developed. Through this interface, users, from clinicians to researchers, can enter patient specific parameters applicable to Warfarin dosing. These comprise demographic attributes (Gender, Race and Age Group), anthropometric measurements (Height (cm) and Weight (kg)) as well as clinical factors (Diabetes status, concurrent use of Simvastatin and Amiodarone, and the patient's Target INR value).

The interface's management element is the model selection dropdown menu, which allows users to select the trained predictive models, for example, Linear Regression, Random Forest, XGBoost, etc. The interface then provides a predicted therapeutic Warfarin dose in mg/week together with an indication of Low or High therapeutic requirement.

This deployment connects machine learning model development to practical real-world applications by providing a convenient real time decision-support tool. It enables healthcare providers to make data-based estimates of Warfarin dosing demand thus possibly minimizing trial and error adjustments while improving the patients' safety.



Warfarin Dose Prediction Tool

Gender: Male

Race: White

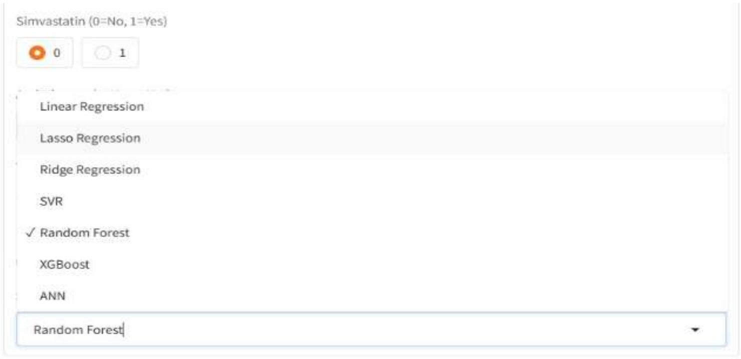
Age Group: 60-69

Height (cm): 193

Weight (kg): 115

Diabetes (0=No, 1=Yes)

Predicted Dose (mg/week): Predicted Dosage: 34.73 mg/week



Simvastatin (0=No, 1=Yes)

● 0 ○ 1

Random Forest

Figure 15: Gradio Interface

Discussion and Conclusion:

The machine learning models developed in this project aimed to predict the therapeutic Warfarin dose based on patient-specific clinical, demographic, and genetic factors. Among all models tested, Linear Regression achieved the best performance with a Test R² score of 0.43, indicating that it was able to explain 43% of the variance in Warfarin dosing. Surprisingly, Linear Regression outperformed more complex models like Random Forest, XGBoost, and ANN. This outcome suggests that the relationships between input features and Warfarin dose in this dataset are primarily linear, with limited benefit from more sophisticated non-linear models given the available features.

XGBoost and ANN were quite competitive, with XGBoost with a Test R² of 0.42, ANN approximately 0.38 showing they can capture non-linear interactions. Their marginal improvement was not worth the extra complexity of linear models, however, in this case. Lasso Regression and SVR were not successful probably because they were too regularized and feature scaling sensitive.

Analysis of features' importance for Random Forest and Lasso Regression indicated Weight, INR Therapeutic, Height, and VKORC1 genotypes to be the most important predictors conform to the pharmacogenetic evidence. These findings confirm the suitability of the model for interpretability and the tuning to clinical understanding.

For deployment, the project resulted in a Gradio-based web application by which users can enter patient data and international normalized ratio (INR) targets and obtain Warfarin dose predictions in real time, representing the usability of machine learning as clinical decision support.

Future Work:

Take on greater and more varied datasets that include continuous variables (such as actual age values, finer race categories).

Experiment with the advanced ensemble techniques like model stacking, for combining strengths of several models.

Embed time-series INR data to simulate dynamic dose adjustments in the periods of therapy initiation and stabilization.

Validate it externally on external cohorts of independent clinical patients as a point to evaluate its generalizability.

Create a cloud-hosted API for use within EHR solutions for practical use.

Contribution:

This project was completed through equal collaboration between Zaheer Khan and Tanisha Londhe. The roles and responsibilities were divided as follows:

Zaheer Khan:

Data acquisition, cleaning, and preprocessing.

Exploratory Data Analysis (EDA) and visualization of feature distributions and correlations.

Implementation of traditional machine learning models (Linear, Ridge, Lasso, Random Forest).

Performance evaluation and preparation of model comparison metrics.

Tanisha Londhe:

Feature engineering, encoding, and data scaling.

Development and tuning of advanced models (XGBoost, SVR, ANN).

Feature importance analysis and visualization (Random Forest, Lasso).

Deployment of models through the Gradio interface and final presentation preparation.

We contributed equally to:

Model selection strategy and experimental design.

Interpretation of results and writing of the final report.

Participation in team meetings and project discussions.

GoogleColablink:

https://colab.research.google.com/drive/1kQUlxej44slEg8jP1zarhsyDIrFBX1bu?usp=sharing

Zoom Recording:

https://slu.zoom.us/rec/share/OQLqrTa5ZfKWaGOWJzfXm2GjJ8ZbAQrHiz11BKF4jWilgZVmvt8EZzAK7K2Z2Iq0.k9CLw9IUMes9W12U

Passcode: x%8lEiTX