ISSN: 2093-5374 / E-ISSN: 2093-5382

Machine Learning Side Effect Trend Predictions and the SIDER Database

Stephen Onorato Jr[1], Md Amiruzzaman[2], Rizal Mohd Nor[3], Md. Rajibul Islam[4], and Dr. Ilsun You[5*]

[1]Department of Computer Science, West Chester University, 700 S High St, West Chester, Pennsylvania, USA. steveonoratojr@gmail.com, https://orcid.org/0009-0007-8786-0538

[2]Department of Computer Science, West Chester University, 700 S High St, West Chester, Pennsylvania, USA. mamiruzzaman@wcupa.edu, https://orcid.org/0000-0002-2292-5798

[3]Department of Computer Science, Kulliyyah of Information and Communication Technology, Kuala Lumpur, Malaysia. rizalmohdnor@iium.edu.my, https://orcid.org/0000-0002-8994-2234

[4]Department of Computer Science and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh. md.rajibul.islam@gmail.com, https://orcid.org/0000-0003-0565-6917

[5*]Department of Information Security, Cryptology, and Mathematics, Kookmin University, Seoul, South Korea. ilsunu@gmail.com, https://orcid.org/0000-0002-0604-3445

Received: July 22, 2024; Revised: August 30, 2024; Accepted: September 30, 2024; Published: December 30, 2024

Abstract

In the Pharmaceutical and Healthcare industries, understanding medications is key in the treatment of patients. Worldwide, there are hundreds of thousands of medications available, classified in categories related to medication therapy and the remediation that they provide. With so many different types of medication, medical doctors and pharmacists need to determine what kinds of drugs to provide to patients with specific medical needs. New medication studies necessitate careful analysis of available medication data during clinical trials, prior to production of new medications, and through the course of prescribed medication therapy. he use of medication therapy is not justified if the number of side effects outweighs the remedial benefits. Therefore, not all medications are deemed medically safe for all patients. Supervised machine learning techniques assist scientists with predicting side effects of medications that are under development. Prediction techniques aid future development of medications based on the properties of current medication data models.

Keywords: Machine Learning, Side Effect Prediction, Pharmaceutical Industry, SIDER Database, Medication Therapy, Clinical Trials, Drug Safety, Supervised Learning

1 Introduction

Medication side effects are unintended reactions a drug may exhibit on the body during medication therapy. While some of these reactions can be positive or neutral, many drugs have undesirable and adverse side effects which can range from mild inconveniences, such as a headache or dry mouth, to

Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA), volume: 15, number: 4 (December), pp. 90-108. DOI: *10.58346/JOWUA.2024.I4.007*

*Corresponding author: Department of Information Security, Cryptology, and Mathematics, Kookmin University, Seoul, South Korea.

life-threatening, such as blood clots and organ damage. When prescribing medications, a healthcare provider must take into consideration the patient's health and if the patient is able to tolerate adverse side effects based on their specific health conditions. A drug that has a significant amount of side effects where the risks outweigh the benefits will be deemed dangerous and potentially fatal if consumed. (FDA, n.d.).

New medications undergo a development process before they are ready for production. During this time, the medication is researched and tested on people and animals in the laboratory, commonly referred to as a clinical trial. During the clinical trial process, scientists measure and record medication efficacy and any side effects exhibited by the test subject. If a clinically trialed medication is deemed safe, the FDA will further review, test, and monitor the drug during production. It could take up to 15 years before a medication passes through the clinical trial process (Cancer Research UK, 2022).

Supervised machine learning is a tool data scientists utilize for training data and expediting the clinical trial process. In pharmaceutical and health studies, scientists employ historical and current medication data models for analysis and computation. By training the data model through supervised inputs and outputs, a scientist can successfully train a dataset to predict a series of well-defined side effect trends. Specifically, medication data models are available for predicting the side effects of drugs undergoing clinical trials. This insight allows data scientists to reduce the time it takes to research and analyze medications for clinical trials, thus expediting the time needed for medication development.

2 Problem Statement and Related Work

The clinical trial period prior to production and manufacturing of new medications can take up to 10 to 15 years to be approved via clinical trial. During clinical trials, humans and animals participate in studies that measure drug efficacy and emergent side effects. Machine learning is used to predict side effects for research medications during the production period. A data set of clinically trialed medications containing their known side effects, in conjunction with medication anatomical, therapeutic, and chemical compositional (ATC) data, is used to form a data model. This model is instrumental in the analysis of frequently appearing side effects of drugs and correlations of side effects with other medication attributes. The model can be trained to predict side effects and ATC trends in future clinical trials.

In the journal "An extensive survey on the use of supervised machine learning techniques in the past two decades for prediction of drug side effects," the study was aimed to analyze a variety of different observations across multiple medication side effect studies (Das & Mazumder, 2023). The study illustrates how a supervised machine learning approach can predict medication side effects with the use of known medication attribute data. Being able to predict side effects can reduce some of the challenges scientists and drug manufacturers face when analyzing and developing new medications. Despite these findings, the study mentions that there are still complexities when using this type of approach to predicting new medication side effects. Only some attribute pairs can successfully be used to identify side effects. Another issue is that not all medications have side effects. Out of the data that was examined, the number of medications without side effects outweighs the number of medications that have side effects. Future studies in this area may use clustering methods in conjunction with supervised machine learning to appropriately classify medications before assigning side effect labels.

In the journal "Machine learning prediction of side effects for drugs in clinical trials," the study aimed to predict unknown side effects in clinical trials using information available in the SIDER and OFFSIDES database (Galeano & Paccanaro, 2022). The Side Effect Resource database or SIDER contains drug-side effect pair data observed in clinical trial studies. The OFFSIDES database contains

aftermarket side effect pair data. The study's authors manufactured the "Geometric Self-Expressive Model" (GSEM), which uses matrices containing medication and side effect data to train the model with the utilization of known drug-side effect pairs of these databases. The study further details how drug-side effect pair data from clinical trial and aftermarket studies can be used to predict trends in future clinical trials and when new medications enter the market.

3 Methodology and Data

The following approach will be taken throughout the course of this study:

1 Clean and preprocess the SIDER database dataset.

2 Clean and preprocess ATC data for merging with SIDER data.

3 Organize the data into two matrices: a Drugs x Drugs matrix (matrix H) and a Side Effects x Side Effects matrix (matrix X)

4 Employ the GSEM model to create matrix Xb such that: HX + XW = Xb, where Xb = Drugs x Side effects + Side effects x Drugs.

5 Identify testing and training sets of well-defined data based on drug-side effect pairs in the Xb matrix and a side effect representation_threshold of 0.05 by using Random Forests (RF), Linear Regression (LR), and Support Vector Machine (SVM) with Python's Scikit-Learn / "Sklearn" library.

6 Train the data by comparing the test set's side effects with results from the training set with each model.

7 Record the number of trained side effects, Random Forest Area Under Receiver Operating Characteristic (AUROC) scores, Linear Regression R-Squared scores, and Support Vector Machine Accuracy scores.

8 Execute and compare precision, recall, and F1 unit tests for each model.

9 Plot the Side effect ratio reporting frequency for each model.

10 Plot and compare AUROC curves, Precision-Recall curves, confusion matrices, AUROC score distributions, and Linear regression scatter plots for each model.

11 Fit the trained data to the Support Vector Machine Classifier Model.

12 Test the dataset in segments with K-Fold Cross-Validation to measure prediction relationships between each fold and each model. Display the results and metrics of each fold, as well as the top 8 important side effect features.

13 Illustrate Anatomical Therapeutic Chemical (ATC) classification relationships of each model's results with the use of box and whisker plots.

The SIDER Side Effect Resource database was converted to a dataset. The database is available from http://sideeffects.embl.de/ (Letunic, n.d.). The database contains over 5,800 unique side effects, over 1,400 unique medications, and over 139,000 drug-side effect pairs. SIDER data is parsed from.gz files into three segments for further analysis: data containing information pertaining to side effect frequency, data illustrating every side effect for each available medication, and an indications data segment that describes valid reasons to use each medication (Dhimmel, n.d.). This study primarily focuses on the side effect data segment, described here as the "se_df" data frame, for analysis.

Each record in the "se_df" data frame contains four columns. The "drugbank_id" column contains the ID of the medication in the DrugBank database. The "drugbank_name" column represents the name of the medication as it appears in the DrugBank database. The "umis_cui_from_meddra" column contains the ID of the medication in the Medical Dictionary for Regulatory Activities. The "side_effect_name" column contains a side effect for each medication record. Medications with multiple side effects contain several or more separate records each indicating a unique side effect. In total, "se_df" contains over 153,000 unique records.

The Anatomical Therapeutic Chemical (ATC) classification system categorizes medications into pharmacological groups (World Health Organization, n.d.). The ATC structure consists of five levels. A unique code is assigned to each medication. The code can be used to identify each of the five ATC levels pertaining to an individual medication or substance. The first level contains fourteen distinct anatomical or pharmacological groups. The second level breaks the medication down further into a therapeutic or pharmacologic group. Third and fourth levels further define chemical, pharmacologic, or therapeutic subgroups for the medication. The fifth level consists of the medication's chemical substance (or medication's name.) Table 1 illustrates ATC classification:

Table 1: Structure and Representation of ATC Codes

Structure of ATC Code	Level Representation
C	**Cardiovascular System (1st level, anatomical main group)**
C10	**Lipid Modifying Agents (2nd level, therapeutic subgroup)**
C10A	**Lipid Modifying Agents, plain (3rd level, pharmacological subgroup)**
C10AA	**HMG CoA reductase inhibitors (4th level, chemical subgroup)**
C10AA05	**Lipitor (5th level, chemical substance)**

A .csv file containing scraped ATC classification information was sourced from Github and utilized for parsing medication side effects to their appropriate ATC classifications (Fabkury, n.d.). This .csv file contained six columns and over 6,900 rows of ATC records. The "atc_code" column contains the ATC code for a particular medication. The "atc_name" column contains the name of the appropriate ATC level or name of the chemical substance. The final four columns "ddd," "uom," "adm_r," and "note" contain information pertaining to the medication's defined daily dose, unit of measurement, administration route, and notes respectively. The information from the final four columns contains mostly null values and is not utilized.

This study primarily focuses on medications and their side effect attributes in conjunction with machine learning models for side effects prediction in clinical trial settings. Each drug-side effect data point is categorical and must be encoded prior to analyzation and algorithmic computation. ATC classification data is utilized for visualizing data performance and results of the study.

Analysis

This study aims to utilize Python, Pandas, and other tools to accomplish the following objectives:

1 Reproduce the "Geometric Self-Expressive Model" (GSEM) and analyze results across various machine learning algorithmic models.

2 Identify a training set of well-defined side effects across three different machine learning models: Random Forests, Linear Regression, and Support Vector Machine.

3 Evaluate and compare each model's accuracy rates in the prediction of side effect observations.

4 Illustrate findings with charts and graphs for each model.

5 Utilize K-Fold Cross-Validation to measure the accuracy of each model's performance across k equal splits.

6 Apply results to Anatomical Therapeutic Chemical (ATC) classifications and illustrate and compare findings using graphs.

7 Detail the economic usefulness of the study's findings.

Parsing SIDER Database Information to Pandas Data Frames

The information available in the SIDER database is freely able for download as .gz files. The information must be parsed as data frames prior to Python utilization and analysis. The resulting "se_df" side effects data frame is utilized in the subsequent preprocessing procedures (Dhimmel, n.d.).

Combining the Side Effects Data Frame with Parsed Anatomic, Therapeutic, Chemical (ATC) Classification Codes

A scraped ATC classification .csv file, (atc.csv,) sourced from Github, suppliments ATC data for parsing to a data frame (Fabkury, n.d.). The side effects data frame and ATC data frame can be combined with a merge operation.

Each medication name from the side effects data frame must exactly match each medication name in the ATC data frame. To accomplish this:

1 A new column is created within both the side effects and ATC data frames. The column is initialized to the lower-cased medication or ATC name of each respective data frame.

2 The ATC data frame may contain some duplicates. These records are cleaned such that there is only one unique ATC name for each ATC code.

3 The indices of each data frame are reset. (This step is pertinent to a successful merge when using Python Pandas.)

4 The data from each data frame is merged to create a new data frame. This data frame consists of the side effect data frame with an additional ATC code column containing the ATC code for each medication.

Combining Resulting Data Frame with ATC Classification Levels

Additional operations must be performed to merge ATC level names with this new data frame:

1 A new data frame is initialized.

2 One by one, new columns are constructed to represent each ATC level of a record in the data frame.

3 One by one, the ATC code for each record is merged with the ATC name corresponding to the exact ATC code match from the ATC data frame.

4 The indices are reset for the next merge operation.

Incorporating the GSEM Model, Learning Similarity Matrices, and Creating a Combined "Drugs x Side Effects" Matrix

The Generalized Self Representation Model (GSEM) can be used in drug side effect predictions (Galeano & Paccanaro, 2022). Here the GSEM model is observed and applied to predict the side effects of medications under development.

The GSEM model integrates drug and side effect information by learning two binary similarity matrices. Similarity matrix H is a matrix consisting of Drugs x Drugs. Similarity matrix W consists of Side Effects x Side effects. The dimensions of each matrix are equal to the number of unique drugs and unique side effects in the data set respectively. The model generates scores for each side effect pair using the equation:

$$Xb = HX + XW \text{ (Equation 1)}$$

To learn W and H, the following objective functions are minimized:

$$min_W \frac{1}{2} \parallel X - XW \parallel_F^2 + \frac{a}{2} \parallel W \parallel_F^2 + b \parallel W \parallel_1 + \sum i \frac{\mu i}{2} \parallel W \parallel_{D,G_i}^2 2 + \gamma Tr(W) \text{ (Equation 2)}$$

$$min_H \frac{1}{2} \parallel X - HX \parallel_F^2 + \frac{c}{2} \parallel H \parallel_F^2 + d \parallel H \parallel_1 + \sum j \frac{aj}{2} \parallel H \parallel_{D,G_i}^2 2 + \gamma Tr(H) \text{ (Equation 3)}$$

Where $\|.\|_F$ denotes the Frobenius norm, and the terms $\frac{1}{2} \parallel X - XW \parallel_F^2$, $\frac{1}{2} \parallel X - HX \parallel_F^2$ represent self-representation, terms $\frac{a}{2} \parallel W \parallel_F^2 + b \parallel W \parallel_1$, $\frac{c}{2} \parallel H \parallel_F^2 + d \parallel H \parallel_1$ represent sparsity, terms $\sum i \frac{\mu i}{2} \parallel W \parallel_{D,G_i}^2$, $\sum j \frac{aj}{2} \parallel H \parallel_{D,G_i}^2$ represent smoothness, and terms $\gamma Tr(W)$, $\gamma Tr(H)$ represent diagonal penalties.

The Xb matrix, illustrated in Figure 1, is built by applying the GSEM model to the dataset. Any non-binary elements in the Xb matrix are converted to binary integers. A heatmap of the Xb matrix can be observed to ensure correct mappings of medications to side effects. The heatmap will also ensure that all elements in Xb are binary. The resulting Xb matrix will be fully encoded and ready for processing.

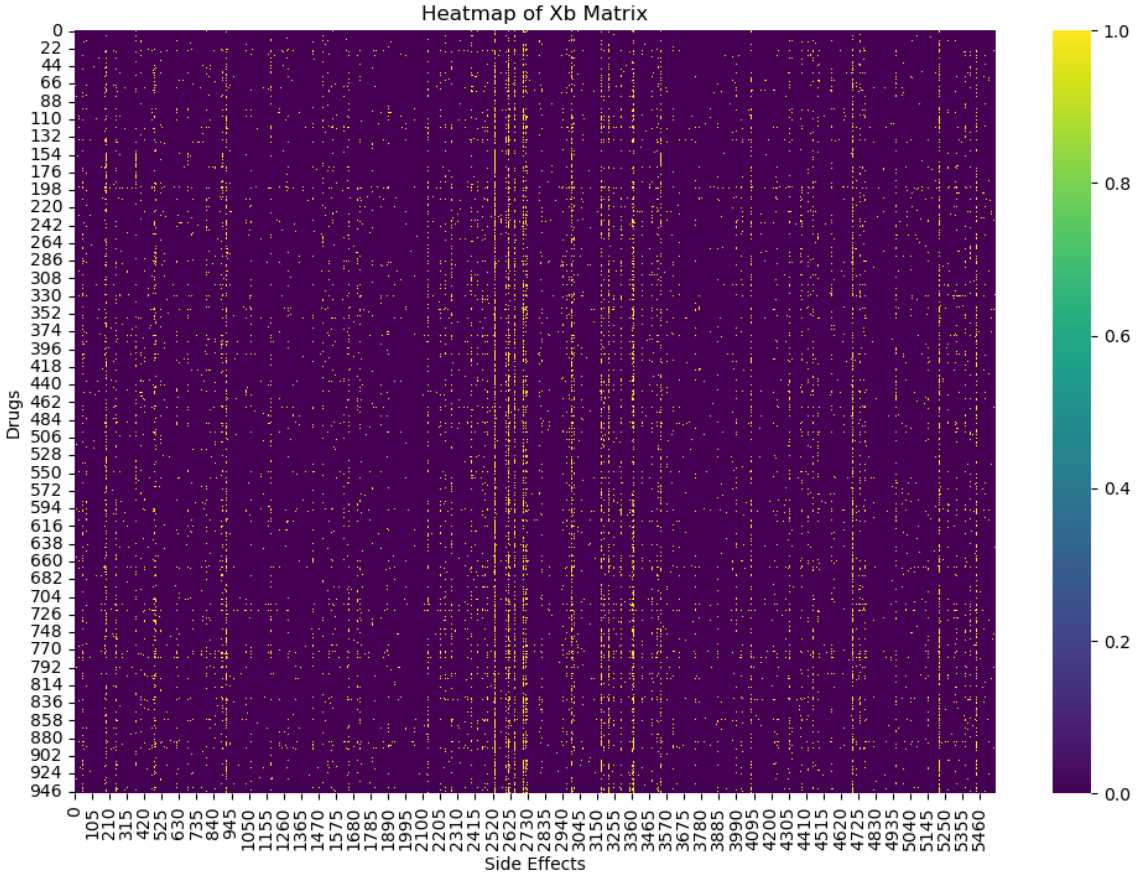

Figure 1: Heatmap of the Xb Matrix Representing Drugs Presenting Side Effects

4 Machine Learning Model Development and Analysis

Scikit-learn combines multiple Python libraries and provides tools for data classification, regression, clustering, dimensionality reduction, model selection, and preprocessing (Learn, n.d.). Scikit-learn's RandomForestClassifier, LinearRegression, and Support Vector Machine (SVC) toolsets are used for machine learning model exploration and evaluation. Additional python libraries, such as matplotlib, seaborn, and numpy, are used in conjunction with Scikit-learn for analyzing and visualizing the results of trained data.

Support Vector Machine is a supervised learning algorithm that can aid in machine learning. It provides ways to analyze data for regression and classification tasks. SVM's primary classification task involves finding an optimal hyperplane between correlating variable data points. A prevalent hyperplane can be used for prediction tasks against trained data.

Linear Regression is a statistical model that is used in regression analysis. Like Support Vector Machine, Linear Regression also analyzes relationships between data points. A linear regression line can be used to show the correlation between attributes and features. This type of analysis could indicate a positive or negative correlation, which can provide insight in machine learning prediction tasks.

The Random Forests model takes an ensemble approach for machine learning tasks. It uses multiple learning algorithms for analyzing data. Training data with Random Forests involves constructing a series of decision trees. Random Forests analyzes and prioritizes results with the greatest number of trees. Statistical calculations can be performed to find the most meaningful or influential data points trained by the model.

Sklearn's "train_test_split()" is a function that extends sklearn's API. It accepts arrays or matrices, (x and y,) and returns a split of testing and training data. Additionally, train_test_split() accepts parameters: test_size, (percentage of the data to be split in the test set,) train_size, (percentage of the data to be split in the train set,) random_state, (a given integer value will result in a repeatable outcome,) shuffle, (a Boolean value that, if true, will shuffle the dataset before preforming the split,) and stratify, (retain class labels for a given matrix or array.)

X is defined as the source feature and y is defined as the target feature, such that: The x variable will contain a list of medications presenting side effects and the y variable will contain a column representing a series of clinically presented side effects. The side effects column is filtered with a threshold to filter out any side effect that isn't presenting by 0.05 (or 5%) of medications.

Testing and training then occurs by looping through each side effect, utilizing Scikit-learn's train_test_split() function on every medication for each side effect. Data is split into 80% training, 10% validation, and 10% testing segments. Test train pairs are then passed to each machine learning model: Random Forest, Linear Regression, and Support Vector Machine.

5 Results

Results for these models are analyzed and compared, including the number of trained side effects, area under receiver operating characteristic (AUROC) scores, r-squared scores, and accuracy scores. Figure 2 illustrates comparisons of Precision, recall, and F1 unit tests for the side effects of diarrhea. Precision indicates the accuracy of positive predictions. Recall indicates the completeness of the predictions. F1 combines both precision and recall into one score. Figure 3 demonstrates recall and precision performance on a precision-recall curve. A high area under the precision-recall curve indicates low false positive rates and low false negative rates for side effect prediction tasks.

```
Random Forest Model Evaluation:
Selected side effect: Diarrhoea
Precision for each class: [1.         0.92405063]
Recall for each class: [0.72727273 1.        ]
F1-score for each class: [0.84210526 0.96052632]

Linear Regression Model Evaluation:
Selected side effect: Diarrhoea
Precision for each class: [1. 1.]
Recall for each class: [1. 1.]
F1-score for each class: [1. 1.]

SVM Model Evaluation:
Selected side effect: Diarrhoea
Precision for each class: [1.         0.96052632]
Recall for each class: [0.86363636 1.        ]
F1-score for each class: [0.92682927 0.97986577]
```

Figure 2: Model Evaluation and Corresponding Precision, Recall, and F1 scores for a Selected Side Effect, Diarrhea

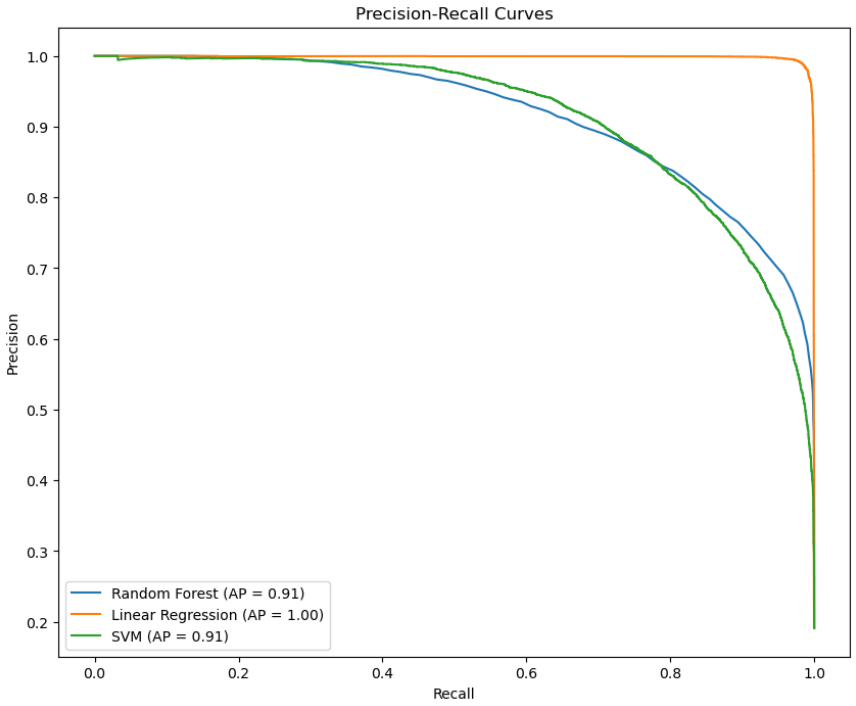

Figure 3: Precision-Recall Curve Representation for RF, LR, and SVM Models

Ratio of Reporting Frequencies

The Ratio of Reporting Frequency (RRF) is a normalized count of medications that are associated with a given side effect. The smaller the RRF, the less medications are associated with a given side effect. Higher RRF represents side effects that are associated with many drugs. RRF was measured and compared between each observed machine learning model. Figures 4 through 6 show the RRF when compared against Random Forest's AUROC scores, Linear Regression's R-squared scores, and Support Vector Machine's accuracy scores respectively.

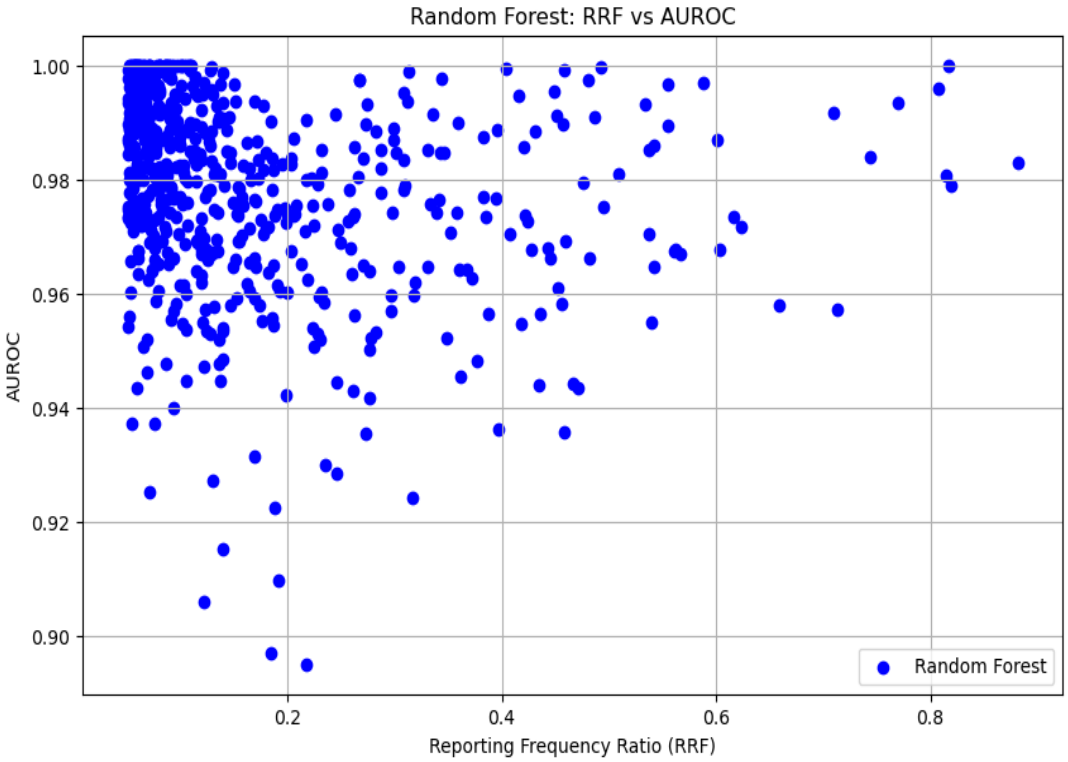

Figure 4: RRF Compared against Random Forest's AUROC Scores

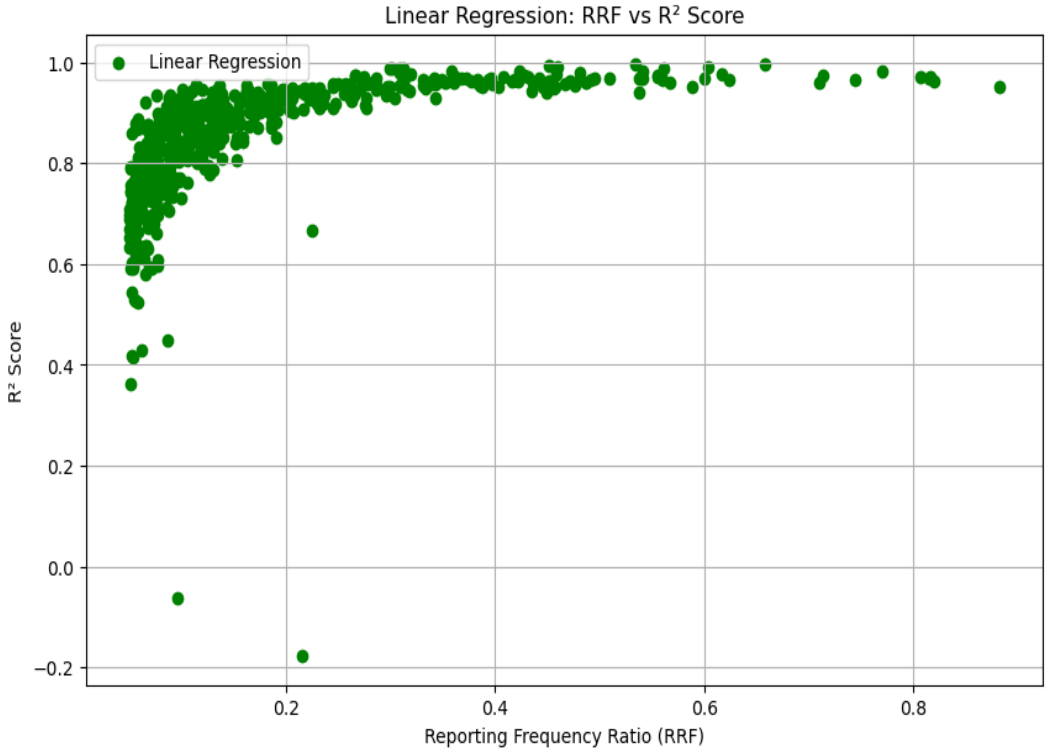

Figure 5: RRF Compared against Linear Regression's R-squared Scores

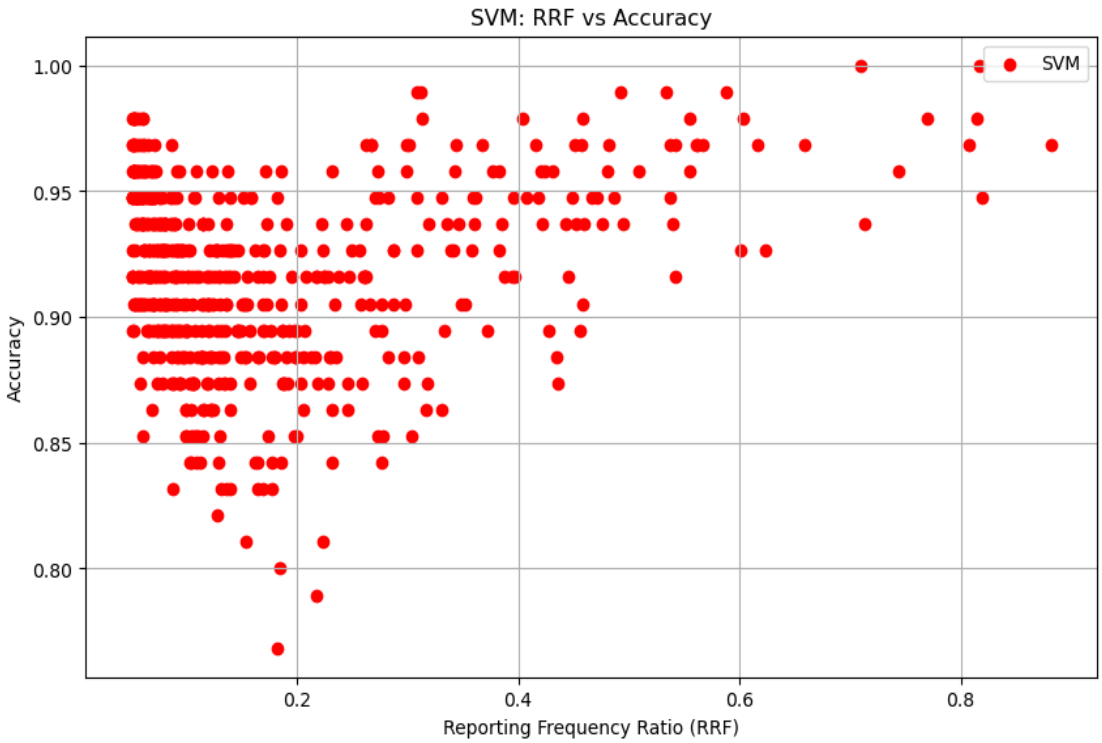

Figure 6: RRF Compared against Support Vector Machine's Accuracy Scores

Receiver Operating Characteristics (ROC) Curves

ROC curves measure and plot the true positive rate against the false positive rate. The true positive rate (TPR) also known as "sensitivity" or "recall" is the ratio of correctly predicted positive observations to actual positive values. The false positive rate (FPR) is the ratio of incorrectly predicted positive observations to actual negatives. Formulas for TPR and FPR are described below:

$$TPR = \frac{True\ Positives}{(True\ Positives + False\ Negatives)} \qquad FPR = \frac{False\ Positives}{(False\ Positives + True\ Negatives)}$$

A ROC curve along the diagonal line with a slope of m = 1 represents a random classifier. Lines along the random classifier represent a trait that is no better than random guessing. Lines above the classifier represent good performance with prediction tasks and coincide with higher TPR for lower FPR. Figure 7 shows ROC curves for Random Forest, Linear Regression, and SVM models for the selected side effect coagulopathy. Figure 8 shows the distribution of AUROC scores for each model across all trained side effects.

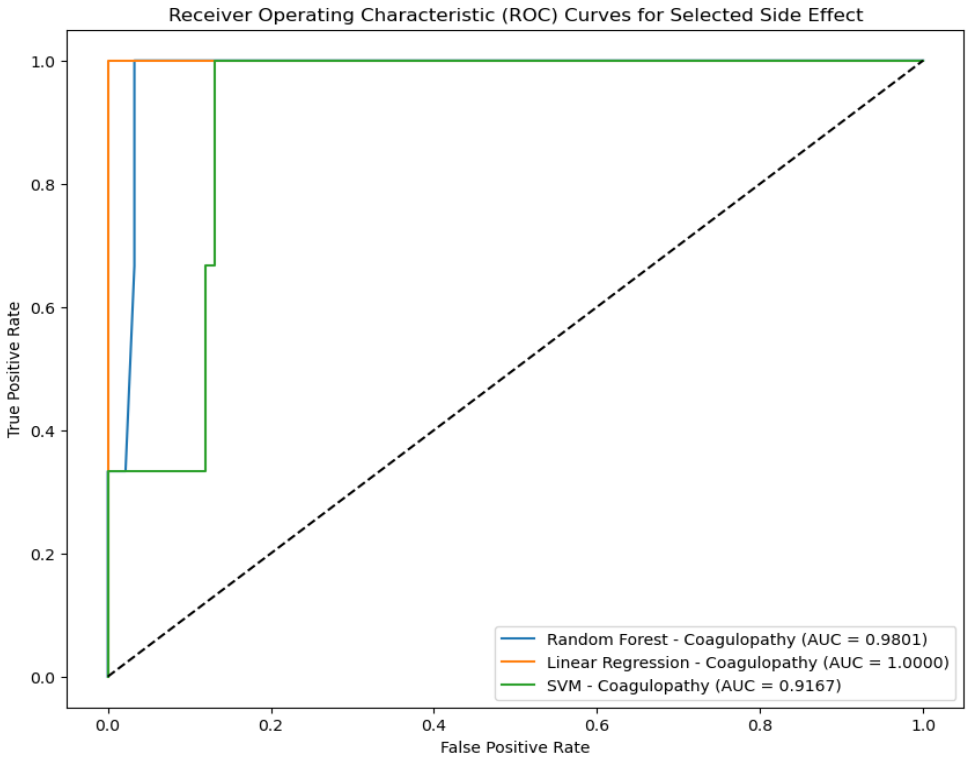

Figure 7: ROC Curves for each Machine Learning Model against the Side Effect, Coagulopathy

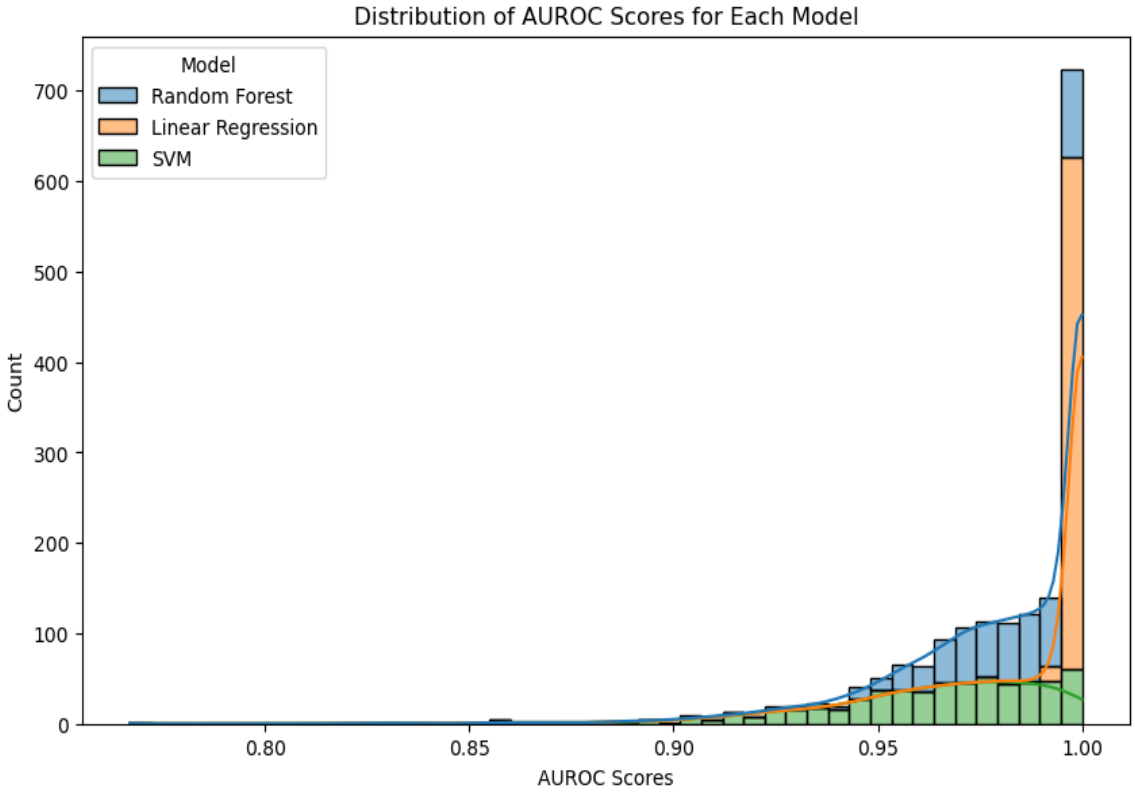

Figure 8: The Distribution of All Side Effect ROC Scores for each Model

Confusion Matrices

A confusion matrix is a table that illustrates a side effect's True Negative, True Positive, False Negative, and False Positive predictions, where:

- True Negative represents the accurate predicted absence rate of a side effect
- True Positive represents the accurate predicted presence of a side effect
- False Negative incorrectly predicts the absence of a side effect when it is present
- False Positive incorrectly predicts the presence of a side effect when it is not present

Table 2 shows a textual representation of a confusion matrix.

Table 2: Structural Representation of a Confusion Matrix

	Predicted Negative (0)	**Predicted Positive (1)**
Actual Negative (0)	**True Negative (TN)**	**False Positive (FP)**
Actual Positive (1)	**False Negative (FN)**	**True Positive (TP)**

Figures 9, 10, and 11 present confusion matrices for the thrombocytopenia side effect. Figure 9 shows a confusion matrix generated with Random Forest model predictions, Figure 10 shows a confusion matrix generated with Linear Regression model predictions, and Figure 12 shows a confusion matrix generated with Support Vector Machine model predictions.

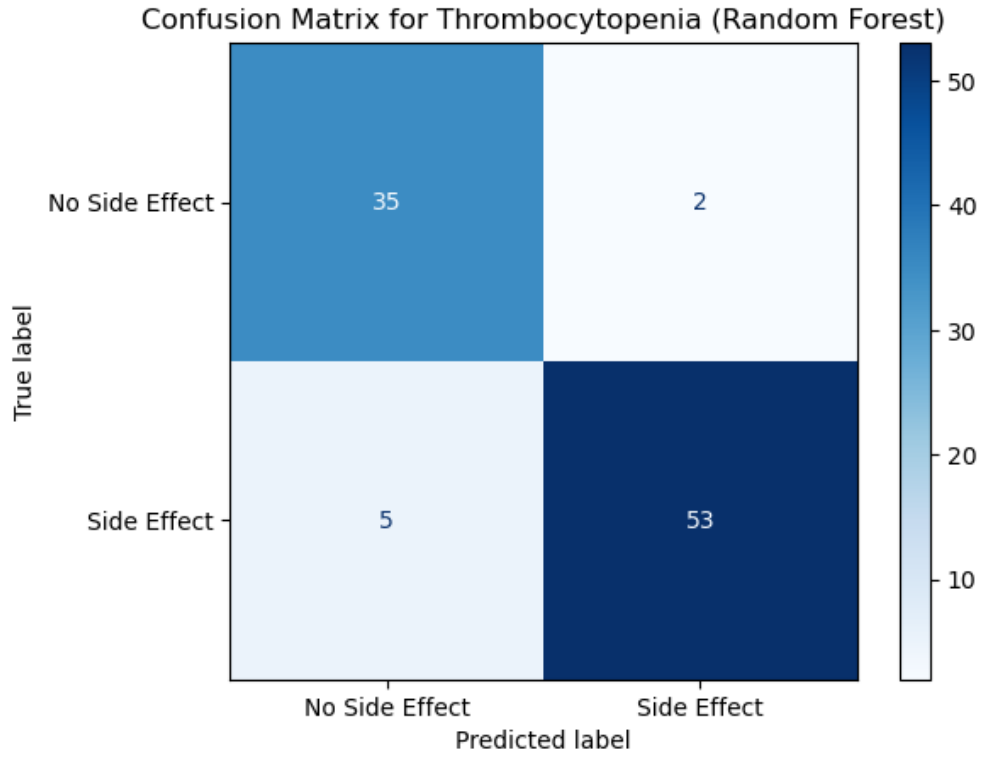

Figure 9: Random Forest Confusion Matrix Predictions for the Thrombocytopenia Side Effect

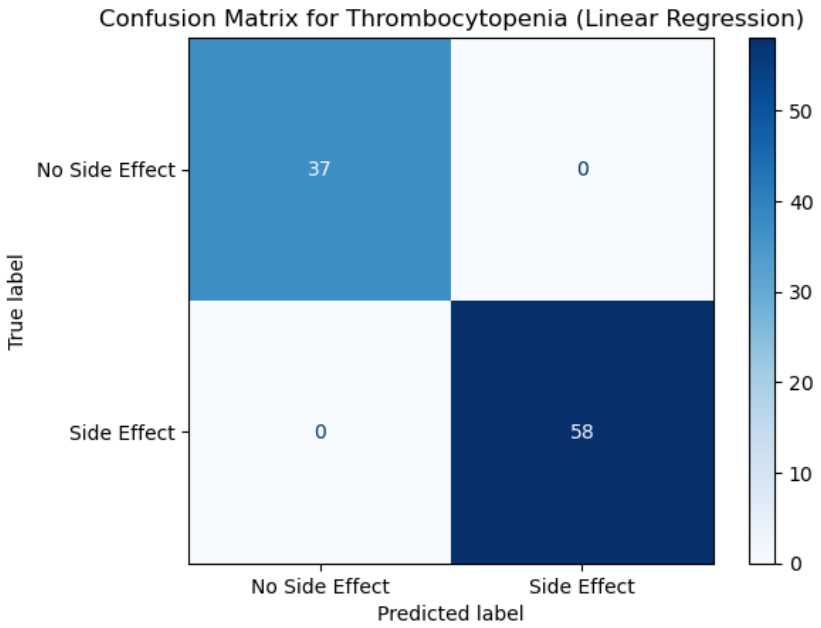

Figure 10: Linear Regression Confusion Matrix Predictions for the Thrombocytopenia Side Effect

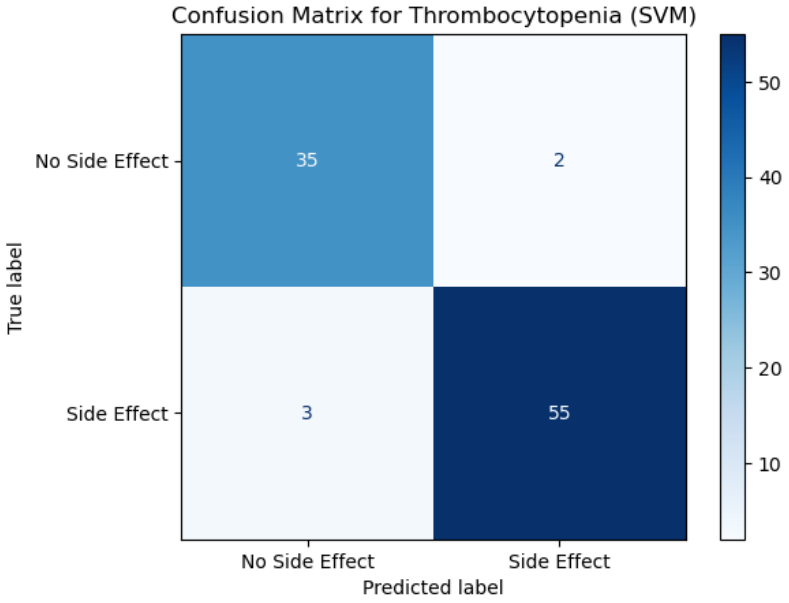

Figure 11: Support Vector Machine Confusion Matrix Predictions for the Thrombocytopenia Side Effect

Linear Regression Scatter Plots

Each data point on the linear regression scatter plot represents either the absence or presence of a side effect. A positive slope indicates positive correlation between actual and predicted values. The closer the linear regression line is to the slope, where m = 1, the more accurate the model is at predicting the absence or presence of side effects. Figure 12 illustrates a linear regression scatter plot for the side effect anaphylactic shock. Figure 13 illustrates a linear regression scatter plot across all side effects.

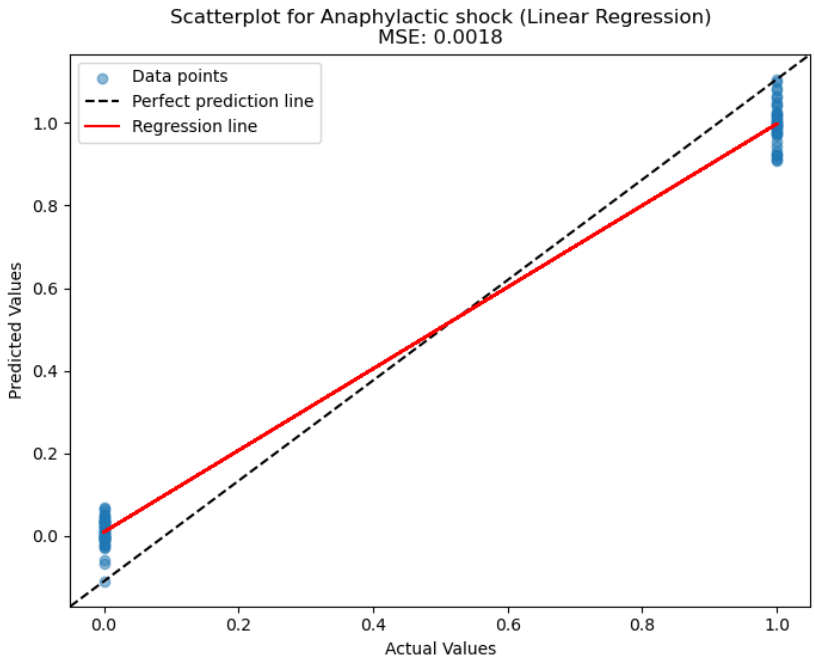

Figure 12: Linear Regression Scatter Plot for Anaphylactic Shock

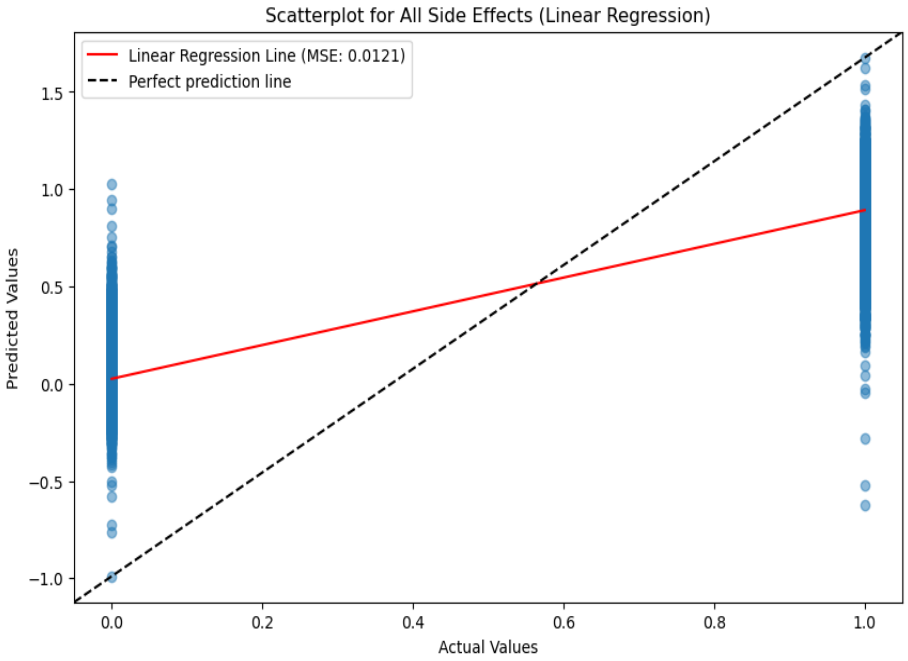

Figure 13: Linear Regression Scatter Plot for All Side Effects

K-fold Cross-Validation Metrics

K-fold cross-validation breaks the data set into k equal segments. Each model uses testing and training parts from each of the k segments in a new testing and training set. This ensures that each model works accurately by testing it against different subsets of the data set. K is set to five separate folds. Each model is looped k times.

```
Random Forest Metrics (k=5):
Accuracy Scores: [0.9315789473684211, 0.9263157894736842, 0.9526315789473684, 0.92105263157
89473, 0.9365079365079365]
Precision Scores: [1.0, 1.0, 1.0, 1.0, 1.0]
Recall Scores: [0.23529411764705882, 0.06666666666666667, 0.18181818181818182, 0.1176470588
2352941, 0.14285714285714285]
F1 Scores: [0.38095238095238093, 0.125, 0.3076923076923077, 0.21052631578947367, 0.25]
Average Accuracy: 0.9336
Average Precision: 1.0000
Average Recall: 0.1489
Average F1 Score: 0.2548

Linear Regression Metrics (k=5):
MAE Scores: [0.06086480509041576, 0.050313377021746056, 0.07029856063358847, 0.058761366540
46726, 0.061207558029276805]
R² Scores: [0.9068468459804764, 0.9309256864527015, 0.8307781185872254, 0.9144750088203011,
0.8902191932631704]
Average MAE: 0.0603
Average R²: 0.8946

SVM Metrics (k=5):
Accuracy Scores: [0.9894736842105263, 1.0, 0.9894736842105263, 0.9894736842105263, 0.989417
9894179894]
Precision Scores: [1.0, 1.0, 1.0, 1.0, 1.0]
Recall Scores: [0.8823529411764706, 1.0, 0.8181818181818182, 0.8823529411764706, 0.85714285
71428571]
F1 Scores: [0.9375, 1.0, 0.9, 0.9375, 0.923076923076923]
Average Accuracy: 0.9916
Average Precision: 1.0000
Average Recall: 0.8880
Average F1 Score: 0.9396
```

Figure 14: Prediction Results for each Model for each Fold

Figure 15: Accuracy, precision, recall, F1, mean absolute error, and r-squared metric evaluation and comparison.

Figure 14 shows a line graph representing prediction results for each model across five folds. Figure 15 presents results for accuracy, precision, recall, f1, mean absolute error, and r-squared metric evaluation for each model across five folds.

Side effect coefficients are observed for each model. For Random Forests, higher values indicate the side effect is influential in prediction evaluation. For Linear Regression, positive coefficients indicate that as the weight of the side effect increases, the stronger the relationship between the side effect and the predicted target value. The reverse is true for negative coefficients. For Support Vector Machine, like Linear Regression, positive and negative coefficients represent the influence of a side effect against the decision-making boundary. Figure 16 shows the eight most prevalent feature coefficients for each model for each fold.

```
Feature Importances/coefficients for Random Forest:
Abdominal cramps: 0.1331
Pallor facial: 0.0063
Epigastric distress: 0.0057
Injection site urticaria: 0.0050
Mucosal pigmentation: 0.0049
Adenitis: 0.0047
Injection site fibrosis: 0.0046
Lymphadenitis: 0.0044

Feature Importances/coefficients for Linear Regression:
Intercept: -0.0000
Abdominal cramps: 0.7845
Abdominal pain: 0.1585
Gastrointestinal pain: -0.1284
Abdominal sepsis: 0.0825
AML progression: -0.0824
Abdominal bloating: 0.0526
Bronchial hyperreactivity: 0.0511
Abdominal aortic aneurysm: -0.0502

Feature Importances/coefficients for SVM:
Intercept: -1.0080
Abdominal cramps: 1.2964
Abdominal pain: 0.4126
Gastrointestinal pain: -0.2241
Nausea: 0.1501
Diarrhoea: 0.1490
Lightheadedness: 0.1054
Epigastric distress: 0.0941
Bronchial hyperreactivity: 0.0941
```

Figure 16: The Eight Most Prevalent Feature Coefficients for each Fold for each Model

Anatomic Therapeutic Chemical (ATC) Classification Box and Whisker Plots

ATC box and whisker plots visualize performance statistics for each model's minimum, first quartile, median, third quartile, and maximum result values. This set of graphs can be utilized to show how accurate machine learning model prediction could be used in clinical trial settings with respect to medications of a particular ATC classification. Results can be applied to all levels of ATC classification. The following graphs demonstrate ATC first level classification results for Random Forests, Linear Regression, and SVM models. Figures 17 through 19 show performance distributions of first level ATC classification across Random Forest, Linear Regression, and Support Vector Machine models respectively.

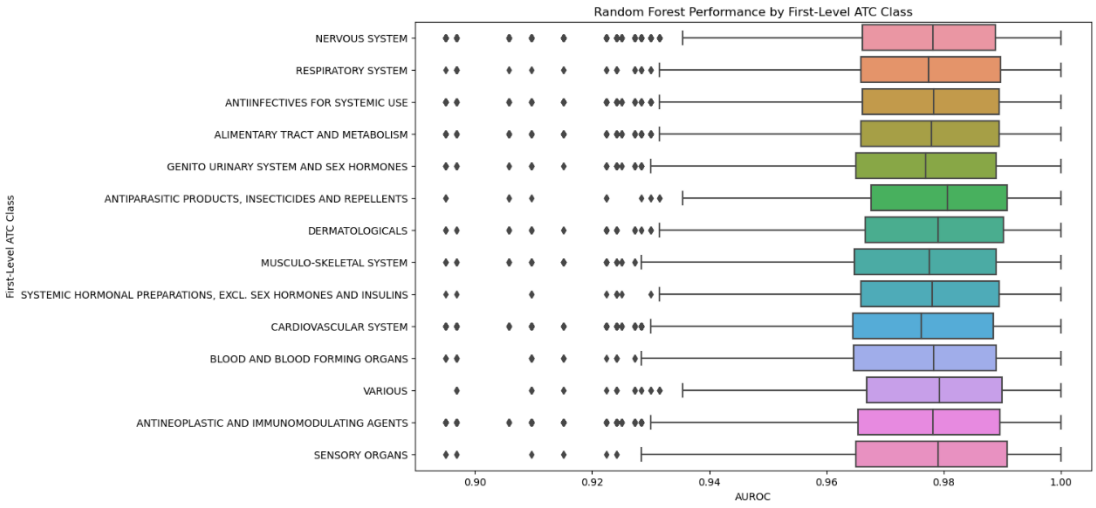

Figure 17: Random Forest Performance by First-level ATC Class

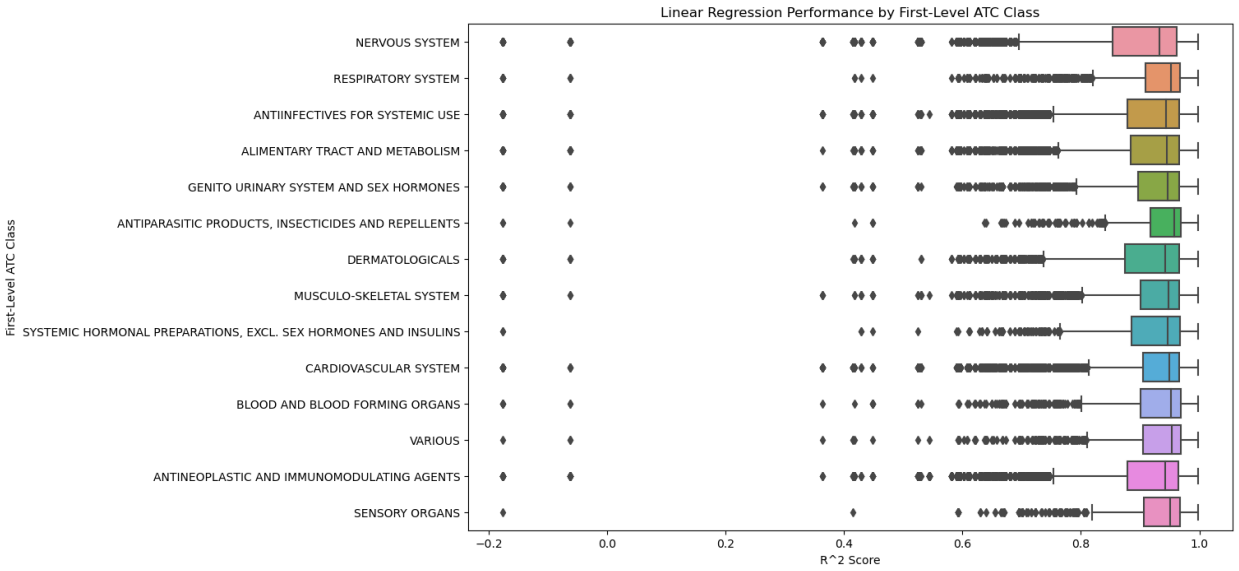

Figure 18: Linear Regression Performance by First-level ATC Class

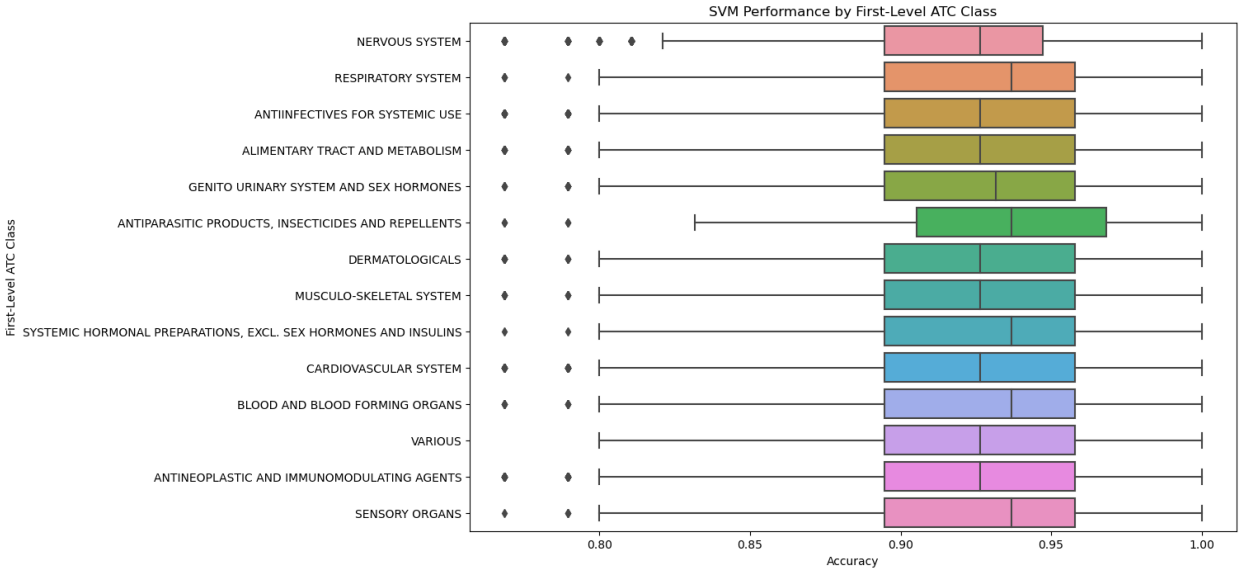

Figure 19: Support Vector Machine Performance by First-level ATC Class

6 Discussion and Conclusion

This study aimed to evaluate data that was available in clinical trial scenarios. Many medications that do not present side effect data are filtered via a threshold. More data would likely improve machine learning model performance.

Sometimes side effects that weren't presented in clinical trial settings will manifest in medication therapy postproduction. Maintainers of the OFFSIDES database record and store side effects presenting in medications currently on the market (Galeano, Paccanaro, 2022). There are upsides and downsides to using this type of data in clinical trials. One potential upside would be having access to more data for model training. However, depending on the drug manufacturer, different additive fillers could be used during medication production. If a patient is sensitive to an additive and exhibits side effects, it may not

necessarily be the medication chemical composition at fault. This could prove to be a downside to using aftermarket data in machine learning analysis.

Medication therapy is integral to the healthcare industry. Healthcare professionals rely on pharmaceuticals to provide patients with adequate care and to assist with medical necessities. The medication development process can take many years before a new medication is ready for production. As an analytical tool, machine learning has the potential to facilitate shorter testing periods with more concise results for scientists and health care professionals.

Sci-kit Learn is a Python library that contains powerful tools for data analysis and machine learning. Sci-kit Learn functions were leveraged throughout the course of this study. Random Forests, Linear Regression, and Support Vector machine Models were utilized in prediction tasks. Resulting metrics from each model were evaluated, compared, and graphed for visual representations. There are a myriad of methods that can be further explored to analyze future prediction tasks with machine learning.

There are still many unexplored medication attribute pairs that could be conducive to predicting side effects. If machine learning can be used to predict side effect trends, scientists can minimize the need for test subjects in medication clinical trials. This can potentially lead to reduced animal and human testing and expediate the clinical trial process which will lead to faster and more efficient medication production. Ultimately, machine learning advances in healthcare and pharmaceuticals has the potential to save money, reduce lab trials on live specimens, conserve resources and, most importantly, lead to better quality of life and even help to save lives.

References

[1] Anatomical therapeutic chemical (ATC) classification, World Health Organization, https://www.who.int/tools/atc-ddd-toolkit/atc-classification/

[2] Center for Drug Evaluation and Research, Artificial Intelligence and machine learning for drug development, U.S. Food and Drug Administration, https://www.fda.gov/science-research/science-and-research-special-topics/artificial-intelligence-and-machine-learning-aiml-drug-development

[3] Center for Drug Evaluation and Research, Learning about side effects (adverse reactions), U.S. Food and Drug Administration, https://www.fda.gov/drugs/find-information-about-drug/finding-and-learning-about-side-effects-adverse-reactions#:~:text=Side%20effects%2C%20also%20known

[4] Center for Drug Evaluation and Research, Using bayesian statistical approaches to advance our ability to evalua, U.S. Food and Drug Administration, https://www.fda.gov/drugs/cder-small-business-industry-assistance-sbia/using-bayesian-statistical-approaches-advance-our-ability-evaluate-drug-products

[5] Das, P., & Mazumder, D. H. (2023). An extensive survey on the use of supervised machine learning techniques in the past two decades for prediction of drug side effects. *Artificial Intelligence Review*, *56*(9), 9809-9836.

[6] Data Society, Data Science's innovation in pharmaceutical clinical trials, https://datasociety.com/data-science-for-pharmaceutical-trials

[7] Dhimmel, GitHub, https://github.com/dhimmel/SIDER4/blob/master/SIDER4.ipynb

[8] Fabkury, GitHub, https://github.com/fabkury/atcd/blob/master/WHO%20ATC-DDD%202021-12-03.csv

[9] Fukuto, K., Takagi, T., & Tian, Y. S. (2021). Predicting the side effects of drugs using matrix factorization on spontaneous reporting database. *Scientific Reports*, *11*(1), 23942.

[10] Galeano, D., & Paccanaro, A. (2022). Machine learning prediction of side effects for drugs in clinical trials. *Cell Reports Methods*, *2*(12), 100358.

[11] Galeano, D., Li, S., Gerstein, M., & Paccanaro, A. (2020). Predicting the frequencies of drug side effects. *Nature communications*, *11*(1), 4575.
[12] How long a new drug takes to go through clinical trials, Cancer Research UK, https://www.cancerresearchuk.org/about-cancer/find-a-clinical-trial/how-clinical-trials-are-planned-and-organised/how-long-it-takes-for-a-new-drug-to-go-through-clinical-trials#:~:text=It%20might%20take%2010%20to
[13] Huang, T., Lin, K. H., Machado-Vieira, R., Soares, J. C., Jiang, X., & Kim, Y. (2023). Explainable drug side effect prediction via biologically informed graph neural network. *medRxiv*.
[14] Jahid, M. J., & Ruan, J. (2013). An ensemble approach for drug side effect prediction. In *2013 IEEE international conference on bioinformatics and biomedicine*, 440-445.
[15] Kennedy, F., Shearsmith, L., Ayres, M., & et al. (2021). Online monitoring of patient self-reported adverse events in early phase clinical trials: Views from patients, clinicians, and trial staff. *Clinical Trials, 18*(2), 168-179.
[16] Kommu, S., & Carter, C. (2023). Adverse drug reactions. StatPearls.
[17] Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2016). The SIDER database of drugs and side effects. *Nucleic acids research*, *44*(D1), D1075-D1079.
[18] Learn, scikit, https://scikit-learn.org/stable
[19] Letunic, I. "Sider 4.1: Side Effect Resource," SIDER Side Effect Resource, http://sideeffects.embl.de
[20] Michael Bihari, M. (2024), Drug classes: Making sense of what medication classifications mean, Verywell Health, https://www.verywellhealth.com/drug-classes-1123991#:~:text=From%20the%20broadest%20perspective%2C%20you,hundred%20classes%20within%20those%20categories
[21] Miller, M. I., Shih, L. C., & Kolachalama, V. B. (2023). Machine learning in clinical trials: A primer with applications to neurology. *Neurotherapeutics*, *20*(4), 1066-1080.
[22] Onitiu, D., Wachter, S., & Mittelstadt, B. (2024). How AI challenges the medical device regulation: patient safety, benefits, and intended uses. *Journal of Law and the Biosciences*, lsae007.
[23] Routray, R., Tetarenko, N., Abu-Assal, C., Mockute, R., Assuncao, B., Chen, H., & Mingle, E. (2020). Application of augmented intelligence for pharmacovigilance case seriousness determination. *Drug Safety*, *43*, 57-66.
[24] The drug development process, U.S. Food and Drug Administration, https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process
[25] TLab, Data-Driven Drug Safety, https://tatonettilab.org/offsides
[26] Tonoyan, L., & Siraki, A. G. (2024). Machine learning in toxicological sciences: opportunities for assessing drug toxicity. *Frontiers in Drug Discovery*, *4*, 1336025.
[27] Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., & Ghassemi, M. (2021). The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, *22*, 1-15.