

Anticipating potential concomitant effects of unfamiliar pharmaceuticals using feature extracts from the Simplified Molecular Input Line Entry System (SMILES)

Adhitya Balaji¹, Mohit Manchella², Wen-Hao Chiang³, Bo Peng³, and Dr. Xia Ning, PhD³

¹Center Grove High School, Greenwood, IN, and ²Carmel High School, Carmel, IN, and ³Department of Computer & Information Science, Indiana University - Purdue University Indianapolis School of Science, Indianapolis, IN

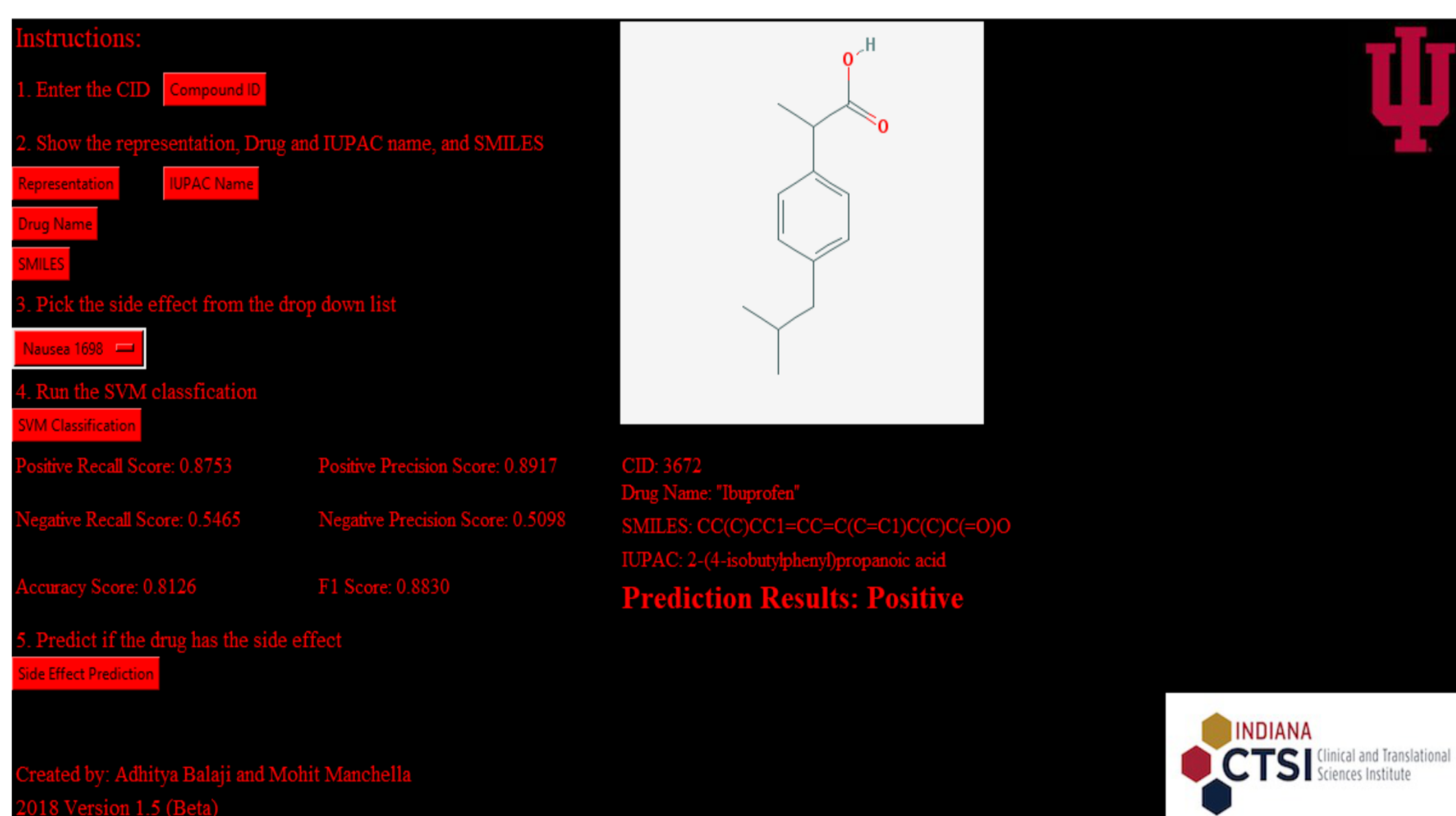
Overview:

There exists a plethora of databases, such as SIDER, that contain information regarding the possible side effects of a number of pharmaceuticals that are available on the market. However, SIDER currently only has side effect data on 1,430 of these pharmaceuticals. Clearly it would be implausible to create a database that would hold the side effect data for every drug available on the market. In some cases there may be pharmaceuticals whose side effects are not well-known as they have not been tested in a clinical setting. In addition, organizations may not have the resources or time to run these experiments which often can be extremely lengthy. A more effective method to collect this information would be to extract the data on the pharmaceuticals that are already well-known and thoroughly tested to create a model that would predict the possible side effects of pharmaceuticals that may not have yet reached clinical trial. By doing so, pharmaceutical organizations can save a plethora of resources by using a model that can accurately predict the side effects of drugs based on their chemical SMILES.

Objective:

To create an application which can accurately predict the concomitant effects of unfamiliar and untested pharmaceuticals using well-studied drugs with similar 2D features and known side effects

Application Window:



Instructions:

1. Enter the CID
2. Show the representation, Drug and IUPAC name, and SMILES
3. Pick the side effect from the drop down list
4. Run the SVM classification
5. Predict if the drug has the side effect

Chemical Structure: CC(C)C1=CC=C(C=C1)C(=O)O

Prediction Results: Positive

Metrics:
Positive Recall Score: 0.8753, Positive Precision Score: 0.8917, Negative Recall Score: 0.5465, Negative Precision Score: 0.5098, Accuracy Score: 0.8126, F1 Score: 0.8830

Drug Information:
CID: 5462328
Drug Name: Diamorphine
SMILES: CC(C)C1=CC=C(C=C1)C(=O)O
IUPAC: 2-(4-isobutylphenyl)propanoic acid

Footer: Created by: Adhitya Balaji and Mohit Manchella, 2018 Version 1.5 (Beta)

Development Methods:

Part 1: In order to be able to predict the side effects of unknown drugs, there has to be a model to use to predict the side effects. This model depends specifically on the side effect it is testing for, so there is no constant model which can be used for every prediction. Therefore, it makes the most sense to create the model and initialize it before the predictions are made. Out of the models we learned, Support Vector Machines (SVM) were the most ideal for our dataset. Unlike a linear regression which creates a straight decision boundary, SVM uses machine learning techniques to seek out and plot the decision boundary to accurately place the value in the respective class (**Figure 1**). Another reason we decided to use a SVM model was because our side effect data was sparse and a linear regression would plot outliers as part of a class, which may not even be accurate. After we created the SVM, we made it automatically run five-fold cross validation and fit the model. After fitting the model, it automatically predicted a value and compared the prediction to the actual and displayed four metrics: precision, recall, accuracy and the F1 score. Accuracy and the F1 score were the most important to our model as accuracy tells us how accurate our overall model is and the F1 score tells us how well the model will do the task we wish.

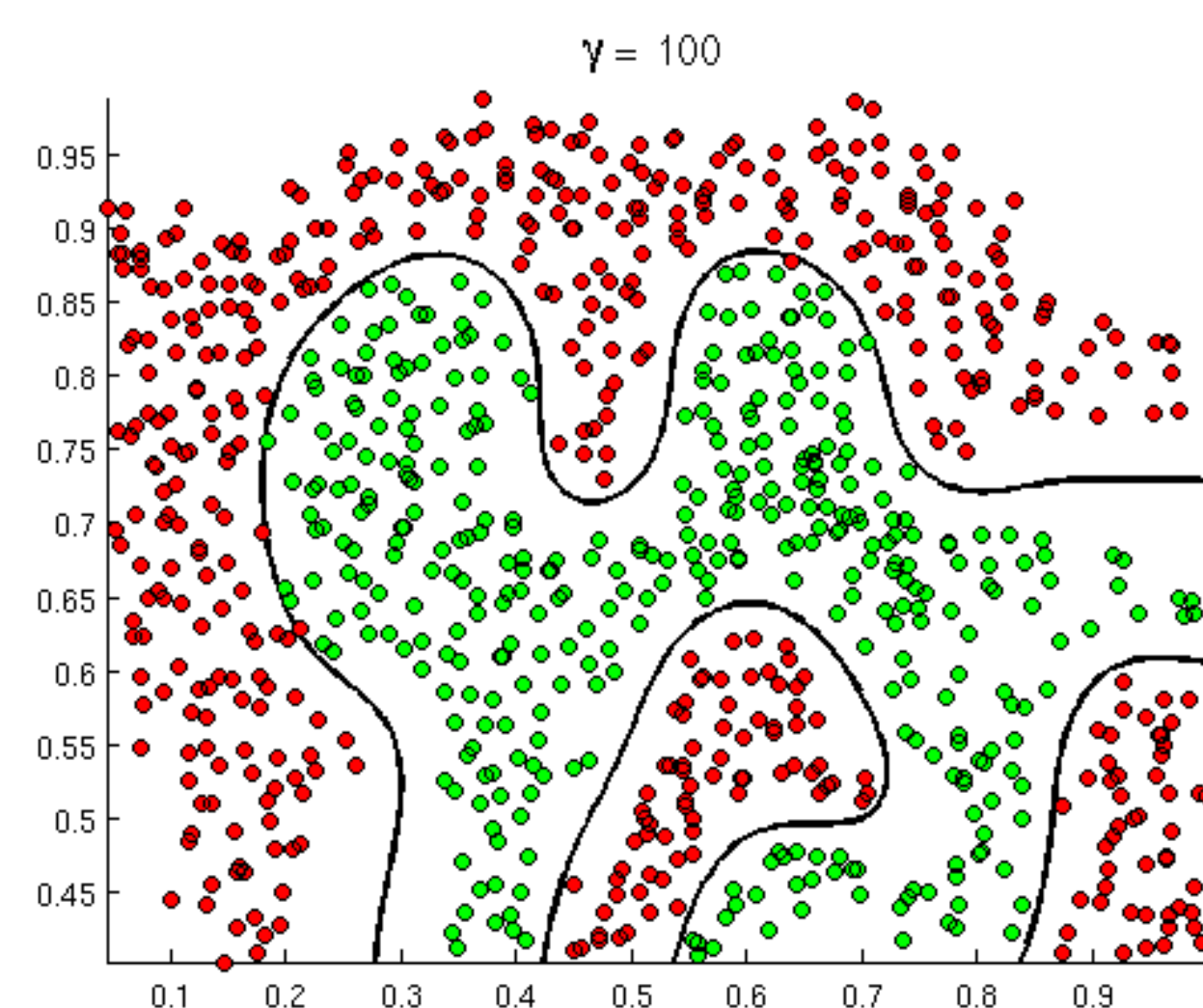


Figure 1: Example of a Support Vector Machine (SVM) with two clearly distinct categories.

Part 2: After creating the SVM, we defined it as a single function and imported it into our GUI. We created buttons which would allow the user to enter a Compound ID (CID) from PubChem, and show the Chemical Representation, the IUPAC Name, and the SMILES. Afterwards, the user can select a side effect from the drop down menu of 5,868 side effects to test for and run the SVM model. Once the model is trained it will display the four metrics mentioned above. Finally, the user can push the side effect prediction button and the program will run the tests and inform the user if the drug is positive or negative for the side effect.

Accuracy Results:

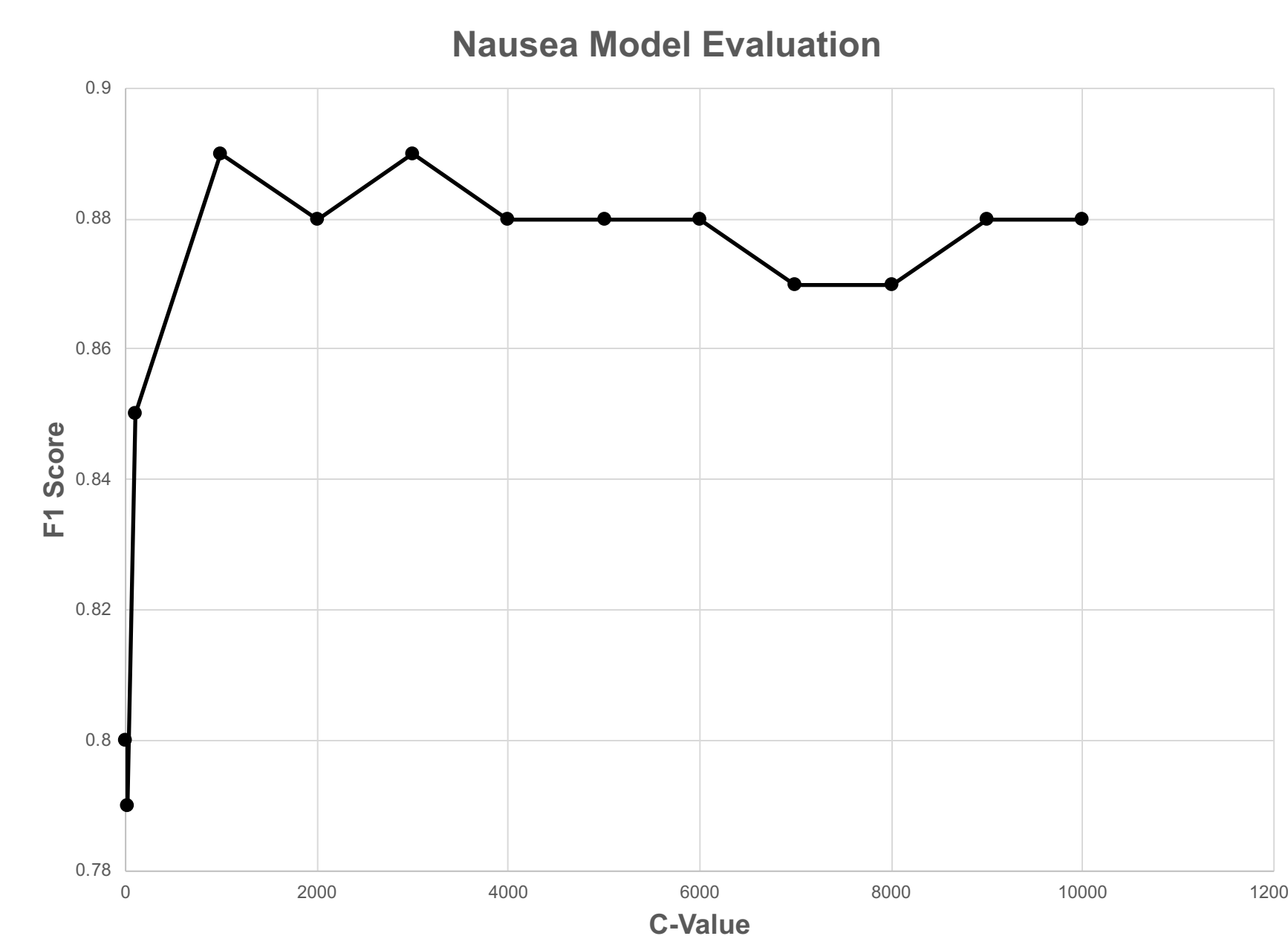


Figure 1: Graphical analysis of the C-values of a one degree polynomial kernel and their respective F1 scores in predicting the *Nausea* side effect.

Positive Drug Name	CID
Diamorphine	5462328
Ceftibuten	5282242
Cefradine	38103
Mepacrine	237
Dibucaine	3025
Alectinib	49806720
Mosapride	119584
Noscapine	275196
Piracetam	4843
Plazomicin	42613186

Figure 2: Known drugs which cause nausea that are not present in the SIDER database, but are predicted positive in our program.

Negative Drug Name	CID
Cocaine	446220
Ondansetron	4595
Promethazine	4927
Phenegan	6014
Prochlorperazine	4917
Dramamine	10660
Meclizine	4034
Hydroxyzine	3658
Cyclizine	6726
Thorazine	2726

Figure 3: Known drugs which do not cause nausea that are not present in the SIDER database, but are predicted negative in our program.

Summary:

Clearly the F1 score can be optimized at a specific C-value. However, the optimum C-value changes depending on the type and degree of the kernel the user wishes to use to make their prediction. Therefore, it is best to leave the choice of parameters up to the discretion of the users as they can appropriately decide what model prediction best suits their needs. Although the overall F1 score and accuracy depend on what C-value the user enters, from our tests we can definitively say that the accuracy will always be above 70%, providing a fairly reliable and accurate result.

Future Improvements:

- Incorporate more machine learning techniques
 - Linear, K-Nearest Neighbors, Decision Trees, Ridge, Lasso, etc.
- Extract different features
 - Use data, such as the compound structural representation, besides SMILES for side effect predictions
- Minor GUI adjustments
 - Make the program more user-friendly and organized

Acknowledgements:

We would like to thank...

- The Indiana CTSI
- Mr. Sanders and the Indianapolis Project SEED/STEM Committee
- Our lab supervisors: Wen-Hao Chiang, Bo Peng
- Our mentor, Dr. Xia Ning
- IUPUI School of Science and the Department of Information and Computer Science
- Our parents

All others who have made this research and application development possible.