Fall 2024 CS/BIOL123A Bioinformatics

Project Report

Project Title: Small Molecule Drug Development for Inhibiting BRAF Protein Mutations Using Machine Learning Algorithms.

Shervan Shahparnia | Data Science | shervan.shahparnia@sjsu.edu
Jason Tobin | Computer Science | jason.tobin@sjsu.edu
Miles Thames | Computer Science | miles.thames@sjsu.edu

Abstract

This report presents the development and application of a Support Vector Machine (SVM) model for predicting the activity of chemical compounds as potential BRAF inhibitors. Using a verified dataset of 303 chemical compounds with activity annotations (“Active” or “Inactive”), rigorous preprocessing steps were performed, including activity mapping, handling missing values, feature scaling, and addressing class imbalance. The dataset was divided into training and testing subsets using an 80-20 stratified split to ensure representative distribution. The SVM model, with an RBF kernel, achieved a prediction accuracy of 97.96% for classifying compounds as BRAF inhibitors, demonstrating its effectiveness compared to traditional linear classifiers. Descriptors for new compounds were computed and used to generate predictions, enabling identification of potentially active compounds with high confidence. This workflow highlights the capability of machine learning to streamline the drug discovery process by accurately classifying compounds based on chemical descriptors. The findings underscore the value of SVM for early-stage drug development, offering a scalable and reproducible approach to identify promising candidates. Future work includes enhancing predictive accuracy through hyperparameter tuning, expanding the dataset to include diverse chemical structures, and introducing advanced metrics such as Expected Predictive Performance (EPP) to assess model reliability. These advancements aim to further optimize the use of computational tools in targeted therapy discovery.

Key Words: BRAF, SVM, Bioinformatics, Machine Learning, BRAF Inhibitors

Table of Contents

List of Figures

INTRODUCTION

BRAF proteins are a part of the MAPK/ERK signaling pathway that is pivotal in communicating chemical signals from the outside of a cell to the nucleus and managing cell proliferation and division. Although proteins are incredibly vital towards the normal operations of the cell, mutations in BRAF are well-known for the possibility of leading to unregulated cell growth, causing multiple forms of cancer. These mutations are found in many different kinds of cancer, including melanoma, colorectal, and thyroid cancer. BRAF mutations are some of the most well-studied causes of unregulated cell growth in cancer research, being “reported to occur in 27.3-87.1% of (thyroid cancers)” (Johansson, Brage). Understanding the implications of these mutations is vital for improvements in current cancer treatment. Just the presence of the mutations alone, can alter the life expectancy and treatment of a patient as they are signifiers of the urgency in which a patient may need treatment. It is important to recognize that due to the frequency of BRAF mutations in multiple kinds of cancers, it is necessary to create individualized treatments for patients due to the variance of frequency and occurrences for different individuals.

However, although BRAF mutations are some of the most common ways cancer can be caused, they can also be prevented. This is done through BRAF inhibitors, a type of small molecule that can stop or slow any possible mutations in the BRAF protein. In current studies, it has been found just how effective BRAF inhibitors actually are when it comes to providing cancer treatment via these molecules. “Targeting BRAF with the use of RAF-selective inhibitors results in remarkable [melanoma] tumor shrinkage” (Villanueva). It is evident that these molecules have good initial testing in current trail treatments.

Since the current scientific theory on inhibitors has proven to be true so far, crucial treatments for patients can begin to be created. Unlike existing cancer treatments, this is completely focused on cancerous only cells rather than both cancerous and healthy cells. This means that it might be possible to see far less side effects in this cancer treatment. Since these molecules have been found to be so incredibly effective in the prevention of cancer causing mutations we chose to attempt to identify these inhibitors. BRAF inhibitors have the possibility to be a major preventative step in providing treatment to patients who could be susceptible to these kinds of cancer. Not only does it provide physicians with a better understanding of the specific treatment a patient needs, but also gives a better quality of life to those being treated. By sparing healthy cells we can seriously make an impact on the prevention and treatment of different types of cancers.

Since BRAF mutations are so common, finding any form of inhibitor can only prove beneficial in multiple areas of oncology; research, treatment, and prevention. Not only might this make treatment easier for both doctors and patients but it could lead to a greater understanding of cancer formation and overall biology all together. It could lead to the development of more effective and targeted cancer treatments, reducing side effects, and improving patient outcomes both during and after the treatment process. By measuring safer and more effective ways of dealing with cancer, it can reduce the amount of invasive treatments that currently are used as forms of treatment.

Although the scope of this project does not necessarily include how we might change cancer research entirely, it is incredibly important to realize the potential in the knowledge gained from studying BRAF mutations in order to see how the field of oncology could change to benefit all affected by any form of cancer.

By improving the current model that exists for the identification of BRAF inhibitors we aim to also improve how cancer is identified and treated. As cancer continues to be one of the driving causes of mortality worldwide, we want to do our part in addressing this issue. Moreover, by focusing specifically on the identification of the small BRAF inhibitor molecules, we are contributing towards a more niche and personalized method of cancer treatment. This means that it is possible to create genetically tailored

treatment plans which can only improve the current area in which people undergo. Not only is this a moral area in which we find it satisfying to pursue this area of research, but authors of this project have direct connections to people who have been affected by the types of cancer mentioned and want to do their part in making a difference towards cancer research and treatment.

We have decided to pursue this area of research through the use of new machine learning libraries and techniques which are constantly evolving in today's world of technology. Machine learning techniques and algorithms can process data magnitudes of order faster than the average human can, allowing us to analyze an incredible amount of data in order to identify patterns and potential small molecule inhibitors. This makes both the discovery process more efficient, but it also makes it more accurate. As long as a model is trained correctly these patterns, and corresponding correlations, become an incredibly powerful tool towards improving existing methods that are commonly used to find these BRAF molecules. Not only does this work with oncology and medicine as a whole, but it is evident that through the use of building a machine learning model, individual treatment plans for patients could be created far easier and treatment could start much faster than before.

This aims to be a very top-level approach towards the identification of BRAF inhibitors. Creating a machine learning model is a singular step in finding new inhibitors and creating a better treatment plan for patients. This method is a cutting-edge approach towards the identification of BRAF inhibitors and will most definitely need even further elaboration to fully utilize the power of machine learning in cancer research.

With this in mind we still aim to analyze our datasets and provide valuable input towards oncology. As mentioned previously, we want to accurately identify the inhibitors, and improve the efficiency and treatment while also improving the personalization of research as a whole. By rapidly screening and speeding up the discovery of these inhibitors it could enable the development of new treatments in a fraction of the time that traditional methods might possibly take.

Ultimately, the goal for the project is creating a world in which cancer is no longer a leading cause of mortality among the human population, however we acknowledge the pure scale of that task and we hope to add our bump to the bubble that is cancer research.

BACKGROUND

In previous work, scientists have used multiple approaches to finding and identifying BRAF inhibitors. One method published in “Identification of BRAF Inhibitors through Silico Screening” decided to attempt to identify these molecules through filtering by solubility and then simulating / predicting how a molecule would bind to a target protein. This method managed to find a series of inhibitors from their database which was about 90,000 compounds in size. They “demonstrate that an efficient and cost-effective virtual screening procedure can be used to identify potent BRAF inhibitors” (Luo). This group found success in their method of filtered data and simulation of bindings.

Our methodologies differ essentially in just the simulation and dataset filtering aspects. In our approach, we are attempting to identify these BRAF inhibitors by building a model using advanced machine learning techniques which predicts the possibility of a molecule being an inhibitor through the chemical composition and structural properties of said molecule alone. This expands the size of the dataset we can use, without the constraint of the solubility of a molecule, as mentioned,

Furthermore, our model does not simulate bindings, rather we are predicting if a certain molecule would bind. This not only means that we can expand the size of the dataset but also that the overall approach is more flexible and can be scaled to a far greater extent. This streamlines the identification process making it possible to find inhibitors faster than through a binding prediction.

Another method of identifying BRAF inhibitors is done in the paper “Identification of a Novel Family of BRAF Inhibitors” where researchers used a high-throughput screening method using an “Elisa-based” system to identify BRAF inhibitors. In the paper, the authors mention that this is a specific detection of BRAF inhibitors. Rather

than as the previous paper mentions using a wide dataset to find many types of inhibitors.

Again, our method differs in the aspect that our model is trained on previously known BRAF inhibitor molecules. This means that, if implemented correctly, the model would be able to identify these specific inhibitors as well as other kinds of inhibitors found in other molecules. This also increases the size of the dataset we are able to comb through as well as increase the total number of molecules found assuming a high accuracy of identification.

DATA COLLECTED / ACCESSED

Our initial plan was to utilize publicly available online datasets from PubChem to gather the chemical information related to BRAF inhibitors. Using RDKit we intended to convert SMILES strings into molecular descriptors and other features that can be used for the model's analysis. This approach was selected to provide as much information possible to the model

The data we need from potential compounds should have key chemical and structural properties that can help distinguish an inhibitor from a non-inhibitor. More data is preferred, but it is important that it contains robust features along with accurate classifications.

While PubChem offers extensive data, tens of thousands of compounds, we could not verify the completeness and accuracy of all the data. Attempting to vet all the data would be time-intensive, and divert resources from the analytical phase.

Our professor foresaw these challenges and provided us with a pre-verified dataset containing real compounds that contained their corresponding molecular features. This dataset eliminated the need for extensive prepossessing and ensured reliability. Our data collection phase was dramatically shortened allowing us to proceed to model tuning.

The dataset was perfect because it provided extensive molecular features that allow the model to understand the chemical properties associated with BRAF inhibition.

The dataset's 'Class' column that indicates whether the compound was an inhibitor or not was essential for our supervised learning approach. The SVM model relies on that label to identify patterns in the descriptors to better classify new compounds.

The dataset contained hundreds of entries with a balanced representation of inhibitors and non inhibitors. This representation ensured that the data would be statistically meaningful and unbiased. The range of compounds and descriptors allows the model to explore many dimensions of the molecular properties of inhibitors. The one concern with the dataset provided is the size. There may not be enough entries to get

APPROACH AND METHOD

The discovery of BRAF inhibitors is a critical area of research in targeted cancer therapies, particularly for conditions such as melanoma and thyroid cancer, where BRAF mutations play a significant role. Traditional drug discovery methods are expensive and time-consuming, requiring extensive vitro and vivo experiments to find potential candidates. As the volume of available chemical data continues to grow, computational approaches are becoming increasingly promising as a way to accelerate the drug discovery process.

This project aims to develop a machine learning based solution to identify BRAF inhibitors efficiently and accurately. By leveraging large datasets of chemical compounds and machine learning techniques, we seek to classify compounds as potential inhibitors or non-inhibitors. This approach is designed to reduce the amount of real-life experiments while providing reliable predictions that can guide further research and development.

Specific Objectives

1. Preprocess and analyze a dataset of chemical compounds with known inhibitory activity against BRAF
2. Design and implement a machine learning pipeline that uses Support Vector Machines (SVM) to classify compounds as inhibitors or non-inhibitors

3. Validate the performance of the model through stratified k-fold-cross-validation and independent test set validation

By achieving these goals this project will contribute to the growing field of AI-driven drug discovery and demonstrates the potential of machine learning to tackle critical challenges in pharmaceutical research.

BRAF's Role in Disease

Mutations in the BRAF gene, such as the BRAF V600E mutation, cause irregular activation of the cell growth signaling pathways, causing uncontrolled cell proliferation. BRAF mutations are implicated in many cancers as mentioned previously, and specifically targeting it with small-molecule inhibitors has been successful for specific cancer treatments. Current inhibitors have their pros and cons as mentioned in the introduction, so there is a constant need for more effective compounds.

Machine learning offers significant advantages in drug discovery. Efficient machine learning algorithms can rapidly analyze thousands of chemical compounds, identifying promising candidates within minutes. Additionally, these models excel at recognizing non linear patterns in chemical features, and inhibitory activity that may not be apparent though conventional methods. Once trained, ML models can scale effectively and predict inhibitors of new untested compounds easily. This computational approach reduces the dependency on expensive and time consuming lab experiments during the early stages. This will not completely eliminate the need for vitro and vivo studies, but will make the ones that occur more efficient. By leveraging these capabilities, machine learning allows researchers to explore vast chemical libraries more efficiently, focus experiments on the most promising candidates, and accelerate the drug discovery process.

APPROACH AND METHODS

Data

As mentioned the dataset used for the project was provided by our Professor which consisted of a csv with chemical compounds labeled as inhibitors or

non-inhibitors of BRAF. Each compound was represented by a set of 356 chemical and molecular features such as molecular weight, heavy atom count, etc.

Preprocessing

This is a crucial step in preparing the data for the model. We started by removing non-numeric columns from the dataset to ensure compatibility with the numerical model. We then checked and handled missing values within the data with the median value of each feature within the column to minimize bias.

We scaled the features in the data. molecular weight and boiling point may have varying scales for example, and the larger number ranges of certain scales could skew the learning process as the larger number could dominate in the distance based calculations. Uneven scaling in a SVM could lead to slower convergence and poorer solutions. Using StandardScaler we made all the features have a mean of 0 and a standard deviation of 1. This ensures that all features of the model contribute equally.

We reduced the dimensionality of the dataset. The larger the dimension (number of features in the data) the relationships between the data becomes harder to detect. Using Principal Component Analysis (PCA) we create new axes that are linear combinations of the original features that maximize the variance in the data. We have our PCA retain 95% of the variance in the data, capturing almost all the critical features, while the noise is discarded. The data is now represented as principal components (axes).

Model Selection

We decided to use the SVM with a Radial Basis Function (RBF) kernel. The SVM finds a hyperplane that best separates the data points belonging to different classes. Although we reduced the dimensional space with PCA there are still many dimensions, which SVM is well equipped to handle. The RBF was chosen because it maps data into an infinite dimensional space allowing it to model complicated patterns that are likely present between BRAF inhibitors.

For the hyperparameters C is the regularization hyperparameter that controls the balance of maximizing the margin, and minimizing the classification errors. The higher the C the smaller the margin, leading to potential overfitting, whereas a small C could allow for too many misclassifications. C=1 was a good balance for this dataset. Gamma

controls the influence of individual data points defining the decision boundary. The larger the gamma, the more close neighbors are considered, creating a more flexible, but potentially overfitted data boundary. A smaller gamma considers far off points that results in a smoother decision boundary. We set the gamma to “scale” which uses scikit-learn’s formula to dynamically adjust the gamma based on the dataset.

This combination of SVM with RBF with our hyperparameters allows for the model to handle and interpret the complex non-linear interactions between molecular features.

Testing and Evaluation

To ensure the reliability and generalizability of our machine learning model, we implement a thorough testing and evaluation pipeline. We included K-fold cross validation, 80-20 test split, and multiple evaluation metrics.

We incorporated stratified k-fold cross validation to generalize and improve the consistency of the model on different subsets of data. Stratification ensures that each fold in the cross validation process keeps the same class distribution as the original dataset. Overall this step reduces the risk of overfitting and underfitting as the model is tested on a variety of data subsets.

We opted for an 80-20 training-testing split to supply a sufficient sample of data for training the model for maximum accuracy. The testing set is independent of the training process, ensuring that the evaluation of the model is reflective of its ability to generalize. We found this balance to deliver excellent results.

We used a variety of evaluation metrics to thoroughly assess the model’s performance. **Accuracy** measures how many predictions were correct out of the total predictions. While it provides a general sense of performance, it can be misleading in cases where the dataset is imbalanced. **Precision**, focuses on how many of the positive predictions (inhibitors) were actually correct. This is particularly important here because misclassifying a non-inhibitor as an inhibitor could lead to wasted resources. **Recall** measures how many of the actual inhibitors were correctly identified by the model. This ensures that the model doesn’t miss potential inhibitors, which is crucial for drug discovery. Finally, we included the **F1-score**, which combines precision and recall into a single metric. By balancing the trade-off between false positives and false negatives,

the F1-score provides a well-rounded view of the model's performance. These metrics together give us a detailed and reliable evaluation of the model.

EVALUATION OF RESULTS

Our original objectives were:

1. Preprocess and analyze a dataset of chemical compounds with known inhibitory activity against BRAF.
2. Design and implement a machine learning pipeline that uses Support Vector Machines (SVM) to classify compounds as inhibitors or non-inhibitors.
3. Validate the performance of the model through stratified k-fold-cross-validation and independent test set validation.

The results of the project demonstrate that the machine learning model successfully met the objectives. The model achieved an overall accuracy of 97.96% on the test set with similarly high performance for individual classes as seen in Figure 2. What is most impressive about the model's results is the quality and correctness of the results. Multiple validation methods were employed, first stratified k-fold cross validation provided a robust evaluation by testing the model on multiple subsets of the data. Our model achieved a mean cross-validation accuracy of 99% with a standard deviation of 1% indicating consistency across folds. Additionally the model validated using an independent test set, which mirrored the cross validation results. This further confirms its ability to generalize effectively. This model is robust and reliable.

A range of metrics were used to evaluate the model. Accuracy was included as a general measure of correctness, which the model scored well in, but is supported by the other stronger metrics. The high precision of 0.98 ensures that the risk of misclassifying non-inhibitors as inhibitors would be minimal. Recall measured the model's ability to not miss a promising inhibitor, which we got a perfect score of 1.00. This guarantees that all potential compounds were identified. The F1 score, being the harmonic mean of precision and recall, provided a balanced metric of overall performance, and we achieved a high score of 0.98.

These results directly answer the key questions of interest. The model demonstrated it can reliably classify compounds as inhibitors and generalize well to unseen data. Both cross-validation and test set evaluations showed consistent results that confirm the model's robustness and applicability.

When initially running the model we were shocked by the initial accuracy metrics. So we went back to implement the evaluation pipeline to ensure that the results were robust. Thorough tuning of the model's hyperparameters and the PCA variance percentage to maximize the model's performance and generalizability.

The confusion matrix (Figure 3) we generated demonstrates that the model has a strong ability to identify inhibitors accurately, and shows that the model is reliable for non-inhibitors. There is slight room for improvement for avoiding false positives.

The choice of evaluation metrics was justified based on the projects' objectives. Accuracy provided a general overview of the performance, precision, recall, and F1-scores were critical to ensure the model's results were reliable. High precision minimizes false positives, reducing the likelihood of wasting resources on non inhibitors. High recall ensured no inhibitors were overlooked, addressing the critical need to identify all potential candidates. The F1-score balanced these metrics, providing a broader evaluation of the model's ability to classify. Together these metrics ensured the model was assessed holistically.

The results confirm that the machine learning model achieved its objectives with the data provided. The results confirm and highlight the potential of machine learning to accelerate and enhance drug discovery efforts. This project provides a strong foundation for future research and application in computational drug discovery.

CONCLUSION AND DISCUSSION

General Observations

As seen in Figure 1 we find that our achieved accuracy was 97.96%, a high value for any machine learning model. This measures how correct the predicted instances from our model turned out to be true. Our model, according to our testing, performs extremely well and is able to accordingly distinguish between BRAF inhibitors and

non-inhibitors. Our precision and recall were also quite high with a value of 1.00 and 0.97. This suggests that for every instance of a non-inhibitor our model was able to predict it every single time. Whereas because the score for the inhibitors were high we can also come to the conclusion that our model was accurate in both the negative and positive aspects.

Furthermore, these metrics collectively imply that the model is reliable not only in predicting true positives but also in minimizing false positives and false negatives. This level of performance is significant for applications in biomedical research, specifically in the exact areas of treatment development that we have been looking into. Our goals throughout this project have been to provide a model that can accurately and precisely determine if a molecule is a BRAF inhibitor and can provide treatment to patients seeking therapeutic and personalized help. Given that we can only go off of what we have so far, we can confidently say that with this dataset we have done just yet. Of course there are more concerns to be looked into with the identified BRAF inhibitors, but by identifying them, our model provides great stepping stones for the scientific community.

Through careful consideration of the data that we have gathered and the results that we have obtained, we are confident in saying that our model for our dataset was a success.

Overfitting Concerns

However, as our dataset was quite small we believe that our model may be overfitting. For the dataset we were given we have come to the conclusion that this is true. Overfitting is when the model learns our dataset too well, this leads to extremely excellent performance but as we are given limited access to our data we can only go off of what we know so far. There are steps to address this which we will cover in the next part.

FUTURE WORK

Model Optimization and Tuning

To enhance the performance of our Support Vector Machine (SVM) model, one possible solution would be to more carefully refine our hyperparameters (like regularization parameters or the kernel coefficient). Methods of accomplishing this could include utilizing Scikit-learn's tools such as Optuna, which automatically tunes hyperparameters. Additionally, we may explore other machine learning models like XGBoost, Random Forest, or some form of Neural Network (FNNs, CNN, RNNs, or LSTMs). These other models may provide better accuracy for prediction.

Data Expansion

Our current dataset, while being reliable and verified by our instructor, contains only 243 individual entries. These entries are described by 357 features that include molecular descriptors such as exact mass, molecular weight, and various physical and chemical properties. The dataset is pre-structured which reduced our time spent studying and preprocessing the data for our model. The limited size of our dataset likely contributes to our very high training accuracy of 97.96%. In the future, to improve our model's learning, we would expand the data using PubChem and DrugBank. Diversifying the data and adding more entries provides a more broad chemical representation and reduces the chances of overfitting, something our model is likely doing due to the limited data.

Improved Evaluation Metrics and Validation

Expanding our dataset would ideally drastically improve our accuracy and other evaluation metrics. However, we may consider using more advanced metrics to validate our model's ability to predict. Using Matthews Correlation Coefficient (MCC) or some form of Receiving Operating Characteristic Curve (ROC) would give better insight into the predictions our model makes by balancing the true and false predictions, but would also likely require a larger dataset to yield reasonable results.

REFERENCES

Hertzman Johansson, C., & Egyhazi Brage, S. (2014). BRAF inhibitors in cancer therapy. In Pharmacology & Therapeutics (Vol. 142, Issue 2, pp. 176–182). Elsevier BV. https://doi.org/10.1016/j.pharmthera.2013.11.011

Luo, C., Xie, P., & Marmorstein, R. (2008). Identification of BRAF Inhibitors through In Silico Screening. In Journal of Medicinal Chemistry (Vol. 51, Issue 19, pp. 6121–6127). American Chemical Society (ACS). https://doi.org/10.1021/jm800539g

Qin, J., Xie, P., Ventocilla, C., Zhou, G., Vultur, A., Chen, Q., Liu, Q., Herlyn, M., Winkler, J., & Marmorstein, R. (2012). Identification of a Novel Family of BRAFV600E Inhibitors. In Journal of Medicinal Chemistry (Vol. 55, Issue 11, pp. 5220–5230). American Chemical Society (ACS). https://doi.org/10.1021/jm3004416

Villanueva, J., Vultur, A., & Herlyn, M. (2011). Resistance to BRAF Inhibitors: Unraveling Mechanisms and Future Treatment Options. In Cancer Research (Vol. 71, Issue 23, pp. 7137–7140). American Association for Cancer Research (AACR). https://doi.org/10.1158/0008-5472.can-11-1243

APPENDICES

Figure 1.

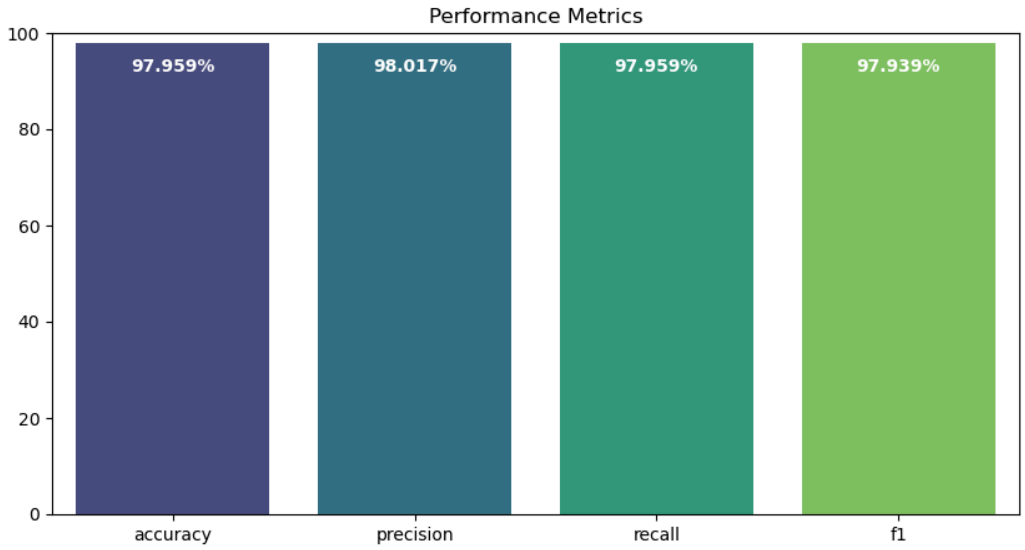


Figure 2.

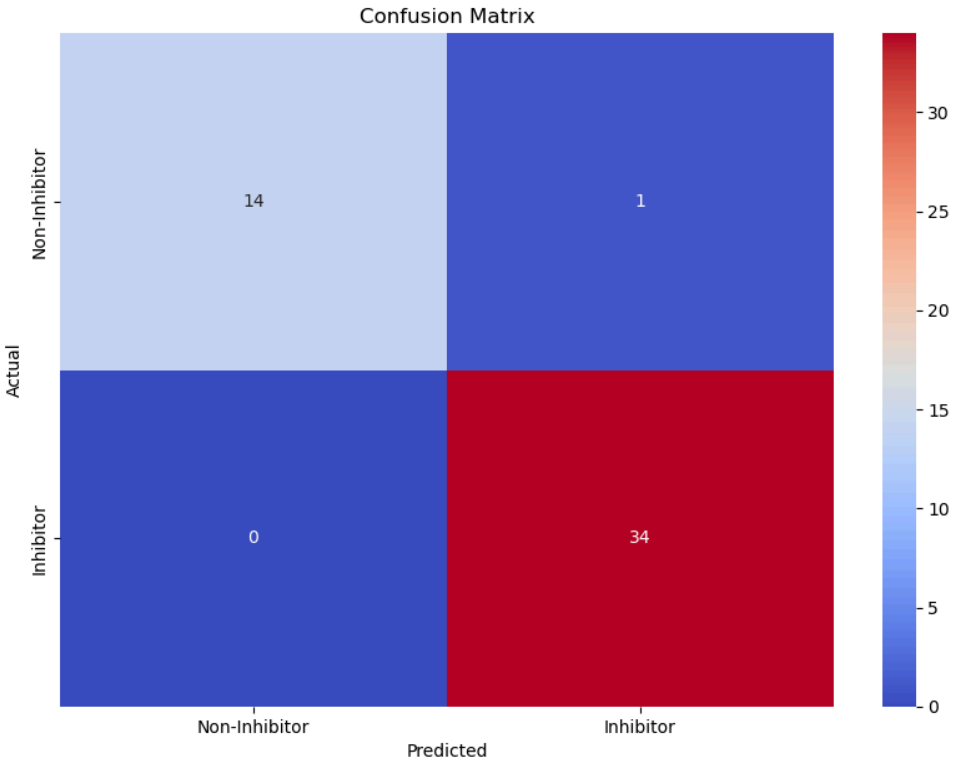
```
Training Complete in 0.21 seconds.
Model Accuracy on Test Set: 97.96%
Model Precision on Test Set: 0.98

Classification Report:
                precision    recall  f1-score   support

Non-Inhibitor       1.00      0.93      0.97        15
    Inhibitor       0.97      1.00      0.99        34

     accuracy                           0.98        49
    macro avg       0.99      0.97      0.98        49
 weighted avg       0.98      0.98      0.98        49
```

Figure 3.


Confusion Matrix
Non-Inhibitor
Inhibitor
Actual
14
1
0
34
Non-Inhibitor
Inhibitor
Predicted
30
25
20
15
10
5
0

CONTRIBUTION OF EACH TEAM MEMBER

Shervan Shahparnia, Jason Tobin, Miles Thames

(1) Total Meetings 5:

All members attended each meeting, except Jason for one who had a family emergency.

(2)

- Jason - Initial and final publication research and model training verification
- Miles - Model training, publication research to back up claims
- Shervan - Model training and code initialization, functions for graphics creation

(3) Areas of the report written:

- Miles - Abstract, approach and method, evaluation of results
- Jason - Introduction, background, conclusion and discussion (with Shervan)
- Shervan - Future results, conclusion and discussion (with Jason)

(4) All members are programmer types so no presentation / powerpoints

(5) Only programmer types so no presentation as well.

(6) Code:

- Shervan - Model initialization, and development
- Miles - 5-fold validation (used to help verify) and model development
- Jason - Confusion matrix generation, and model accuracy checks