4-28-2023

Developing a Machine Learning Model for Protein Conformational Selection and Prediction in Drug Discovery Applications

Eliza Lalitha Asani

Developing a Machine Learning Model for Protein Conformational Selection and Prediction in Drug Discovery Applications

by

Eliza Lalitha Asani

An Honors Capstone
submitted in partial fulfillment of the requirements
for the Honors Diploma
to

The Honors College

of

The University of Alabama in Huntsville

4/30/2023

Honors Capstone Director: Dr. Vineetha Menon
Associate Professor of Computer Science

Eliza Asani (Apr 28, 2023 13:58 CDT) — Apr 28, 2023
Student (signature) — Date

Vineetha — 04/28/2023
Director (signature) — Date

Letha Etzkorn — Digitally signed by Letha Etzkorn Date: 2023.04.28 13:47:26 -05'00'
Department Chair (signature) — Date

4.30.2023
Honors College Dean (signature) — Date

Honors College
Frank Franz Hall
+1 (256) 824-6450 (voice)
+1 (256) 824-7339 (fax)
honors@uah.edu

Honors Thesis Copyright Permission

This form must be signed by the student and submitted as a bound part of the thesis.

In presenting this thesis in partial fulfillment of the requirements for Honors Diploma or Certificate from The University of Alabama in Huntsville, I agree that the Library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by my advisor or, in his/her absence, by the Chair of the Department, Director of the Program, or the Dean of the Honors College. It is also understood that due recognition shall be given to me and to The University of Alabama in Huntsville in any scholarly use which may be made of any material in this thesis.

Eliza Asani

Student Name (printed)

Eliza Asani (Apr 28, 2023 13:58 CDT)

Student Signature

Apr 28, 2023

Date

TABLE OF CONTENTS

Abstract

Computational drug discovery is an important tool that opens the doors for significant advancements in pre-screening for potential drug candidates. One of the first steps in this process requires sampling of protein conformations to test against potential drug candidates for binding affinity. An efficient way to select a sample of conformations would be to classify those conformations as generally binding or nonbinding and to select a representative sample from each of those classes. This project attempted to develop a machine learning model for classifying protein conformations as binding or non-binding using dimensionality reduction and supervised learning techniques. The models were tested on conformations of a single protein and evaluated using performance metrics generated from the confusion matrices. After evaluation the recommended model was the Gaussian Naive Bayes classifier on the original dataset with 30% training data.

Introduction

Computational drug discovery is becoming an increasingly interesting and useful tool for large-scale and varied pre-screening for drug candidates. The basics of computational drug discovery involve modeling the binding of a drug candidate, typically a small molecule called a ligand, with various relevant proteins. Better binding scores indicate a true drug candidate while ligands that do not bind can be filtered out. This pre-screening technique can help significantly reduce the cost in time, labor, and resources associated with pre-screening in a wet lab while also increasing the pool of ligands that can be feasibly explored [1, 2].

Initially, computational drug discovery followed the "lock and key" theory of binding, in which both the protein and ligand are kept rigid. This theory has been replaced by dynamic binding, in which the protein and/or ligand are allowed flexibility to form conformations and move in space. This has come with the understanding that proteins are constantly changing shape, and some of these conformations are much more conducive to binding than others [1].

The application of this binding theory is most represented in ensemble docking, where protein conformations are determined using molecular dynamics simulations and a representative subset of those conformations is then docked against a ligand to determine its binding affinity. A challenge with this method is determining what constitutes a representative or useful sample of protein conformations. One method is to assign a binding affinity, 1 or 0, to each conformation based on its ability to bind to other common ligands. A selection of binding and nonbinding conformations can then be used as a representative sample, perhaps with a bias towards binding in order to promote the discovery of possible drug candidates. The limitation of this method is that it requires computationally expensive docking calculations for each conformation on many different ligands [2].

This project attempts to pursue an alternative method for classifying protein conformations as binding or nonbinding. Starting with an already-classified set of conformations on the human protein ADORA2A, various dimensionality reduction and supervised learning techniques are used to attempt to develop a machine learning model that can accurately classify the binding

affinity of conformations, either binding (1) or nonbinding (0), based on just the protein attributes and without further docking calculations needed. The effectiveness of the model is evaluated using confusion matrix analysis, and the potential strengths and limitations are discussed.

Background and Methods

Dataset

The protein used in this project is the ADORA2A protein, which is a (G-protein)-coupled receptor (GPCR) involved in signal transduction pathways in the human body. GPCRs are transmembrane proteins that receive extracellular signals, in this case adenosine, which are then translated into intracellular signals. ADORA2A is ubiquitous in its effect on the human body as it is involved in processes such as blood circulation and immune function [3].

The ADORA2A dataset utilized in this project contains 3000 protein conformations each with descriptors and binding information. The protein conformations were previously generated from even sampling of a 600 ns coarse grained molecular dynamics simulation on the ADORA2A in its transmembrane environment. Binding affinity - either 0 (nonbinding) or 1 (binding) - was determined by statistical evaluation of docking calculations between ADORA2A conformations and a number of known ligands, as described in [4]. Out of the 3000 conformations 850 are labeled as binding and 2150 are labeled as nonbinding. Finally, 49 descriptors for each conformation were calculated using MOE software, as described in [2]. These descriptors measure properties of each conformation, e.g. hydrophobic surface area, dipole moment, etc., that both distinguish conformations and have an effect on binding affinity. The complete list of 49 descriptors calculated is listed in **Table A.1** in the Appendix.

Dimensionality Reduction

An important component of data preprocessing is dimensionality reduction, which takes n-dimensional data and projects into k-dimensions, where $k < n$. The primary goal of

dimensionality reduction is to reduce the time-complexity of machine learning algorithms on high-dimensional data while still retaining information contained within higher dimensions. Dimensionality reduction can also help guide learning by filtering out less important features and directing learning towards more significant information contained in the data [5]. This project utilizes two common dimensionality reduction techniques: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)

Principal Component Analysis (PCA)

PCA is an unsupervised dimensionality reduction technique that projects n-dimensional data into k uncorrelated dimensions, where $k < n$. The new dimensions are orthogonal to each other and represent the directions of maximum variance as computed from the eigenvalues and eigenvectors of the covariance matrix [5].

Linear Discriminant Analysis (LDA)

LDA is a supervised dimensionality reduction technique that takes into account class-wise separability. If a dataset has C classes, then LDA will reduce the dataset into at most C - 1 dimensions. These dimensions are computed by maximizing the difference in means between classes and minimizing the scatter of each class [5].

Supervised Learning

Machine learning algorithms aimed at classification use supervised learning techniques which require prior knowledge of classes and already-classified data to train on. Supervised learning algorithms split data into a training set and a testing set. The training set is used to fit a model to class-labeled data, and the model is then tested against the testing set with class labels removed. The predicted classification is compared to the actual classification of the testing set in order to determine the strength of the model [6]. This project utilizes three different supervised learning techniques: Support Vector Machines (SVMs), Gaussian Naive Bayes (GNB), and logistic regression.

Support Vector Machines (SVMs)

SVMs separate classes by selecting a hyperplane that represents a decision boundary between classes. There are two main steps conducted by SVMs: representing the data in higher dimensions using a kernel, and selecting a hyperplane. In the first step a kernel function must be chosen within a given set of pre-selected kernels. This project uses two kernels with an SVM classifier: polynomial kernel (Eq. 1), and Gaussian Radial Basis Function (RBF) kernel (Eq. 2).

$$k(x, x') = (scale * \langle x, x' \rangle + offset)degree \tag{1}$$

$$k(x, x') = exp(-\sigma||x - x'||^2) \tag{2}$$

These kernels don't actually transform the data but instead represent the relationship between data points in a relevant dimension without significant computational cost. Once a kernel function is determined, the SVM selects a hyperplane by maximizing a margin between classes, and any data points on the edge or within the selected margin are listed as the support vectors [7].

Gaussian Naive Bayes (GNB)

Naive Bayes (NB) classification is based on Bayes theorem which estimates posterior probabilities. Bayes theorem states that the posterior probability $P(y|x)$ is equal to the product of the prior probability $P(y)$ and the likelihood $P(x|y)$. Assuming the inputs/features are conditionally independent gives a formula for determining the probability of a classification given a set of inputs (Eq. 3). Classification is then conducted by selecting the class with the maximum $P(y|x)$ [2, 8].

$$P(y|x) = P(y)P(x|y) = P(y)\prod_{i=1}^{k} P(x_i|y) \tag{3}$$

Gaussian NB follows the same classification principle but assumes that the inputs are continuous and follow a Gaussian distribution. This results in a formula for likelihood given the mean and standard deviation of the class (Eq. 4) [2].

$$P(x|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp(\frac{-(x-\mu_y)^2}{2\sigma_y^2}) \quad (4)$$

Logistic Regression

Logistic regression is a technique used for binary classification. It calculates a value p which represents the probability of the sample being in class 1. The formula for calculating p is given below, where 1 - p is the probability of the sample being class 0, and the natural log of the ratio of p to 1 - p is equal to a linear combination of k attributes $x_1, x_2, \ldots, x_k$, with weights $\beta_1, \beta_2, \ldots, \beta_k$, and bias β_0.

$$ln(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k \quad (5)$$

The weights and bias are determined iteratively in such a way that minimizes the difference between the actual and predicted classification values. In addition, the threshold percentage that distinguishes classification between the two classes can also be manipulated in order to improve the model [9].

Confusion Matrix Analysis

Confusion matrices can be used to analyze the performance of supervised learning techniques. These matrices compare the actual classifications with predicted classifications from the supervised learning model on the testing set of data. The main diagonals represent the correct classifications while the off-diagonals represent the incorrect classifications.

		Predicted Class	
		0	1
True Class	0	True Negative	False Positive
	1	False Negative	True Positive

Figure 1. Confusion matrix template for binary classification.

The following performance metrics can be calculated from the confusion matrix [2]:

- Accuracy: the number of correctly predicted samples out of the total number of samples
- Sensitivity: the number of correctly predicted positive outcomes out of the total number of positive samples
- Selectivity: the number of correctly predicted negative outcomes out of the total number of negative samples

Results and Discussion

The first step in developing this machine learning model involved pre-processing on the ADORA2A dataset. The 49 relevant features were retained and a min-max scaler was applied in order to even out the effect of each feature on the model. In addition, two dimensionality reduction techniques, PCA and LDA, were applied to see if the data could be effectively reduced into lower dimensions.

Dimensionality Reduction

The explained variance of the principal components on the ADORA2A dataset are plotted in **Figure 2**. While the contribution of each component does visibly decline, the initial contribution of the first principal component is less than 25% and retaining 85% of the variance in the original dataset would require 12 principal components to be used. Therefore, further mention of PCA-reduced data in this work uses 12 principal components.

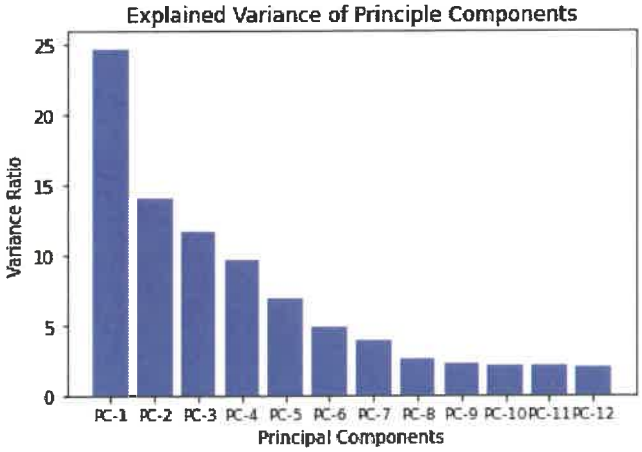

Figure 2. Explained variance of principal components of ADORA2A protein dataset.

In order to further visualize the effect of PCA on the dataset, the data was projected onto the first and second principal components (**Figure 3a**) and the second and third components (**Figure 3b**). It is clear from both plots that there is significant overlap between the binding and nonbinding classes, which suggest that the data is not linearly separable and presents a pessimistic outlook for the classification using the PCA-reduced data.

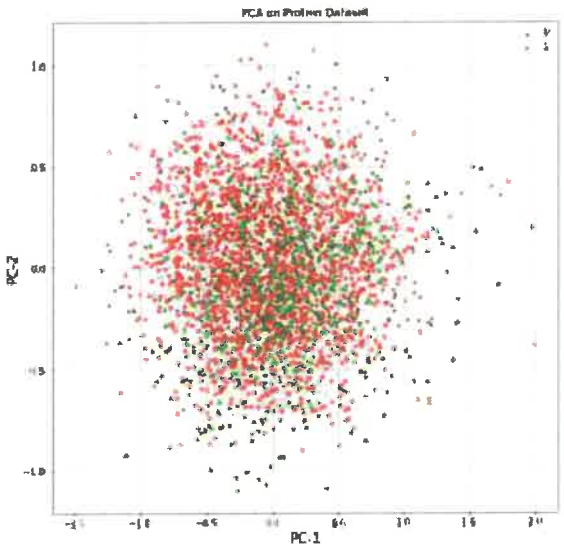

Figure 3a. PCA-transformed data projected onto the first two principal components.

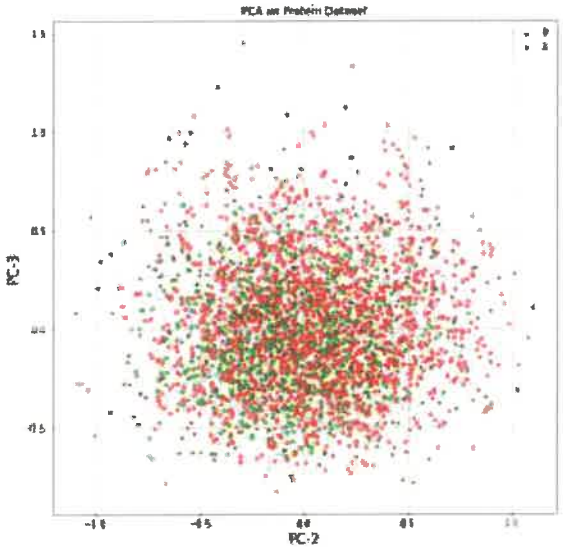

Figure 3b. PCA-transformed data projected onto the second two principal components.

LDA was then performed on the dataset and, given that there are only two classes, the resulting data was projected onto just one dimension (**Figure 4**). While there does seem to be a slight separation between the centers of each of the classes in the LDA-reduced dataset, there is still significant overlap just as with the PCA-reduced dataset.

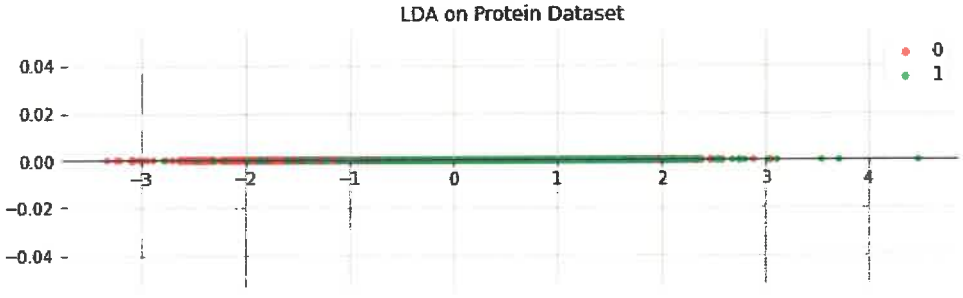

Figure 4. LDA-transformed ADORA2A data projected onto one dimension.

Supervised Learning

Following the data pre-processing, the four supervised learning techniques were performed on the original, PCA-reduced, and LDA-reduced datasets: SVM with RBF kernel, SVM with polynomial kernel, GNB, and logistic regression. Training was conducted on 10%, 20%, and 30% of the dataset, and the cross-validation accuracies and training accuracies were plotted for each method, each training size, and on each dataset (**Figures 5, 6,** 7). The general range for cross-validation accuracy was between 60-75%, with LDA-reduced data being the most consistent between the different supervised learning techniques. Logistic regression and SVM-RBF performed the best over all three datasets. .

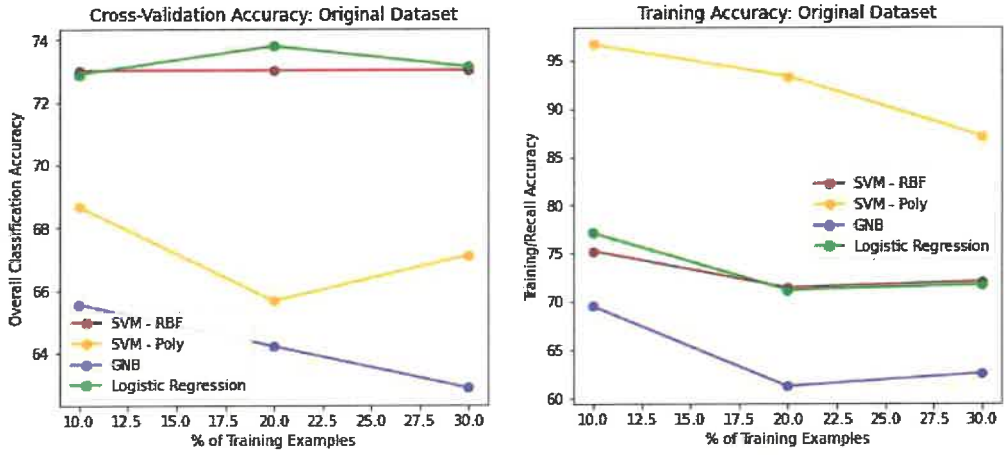

Figure 5. Cross-validation and training accuracy against training size on original dataset.

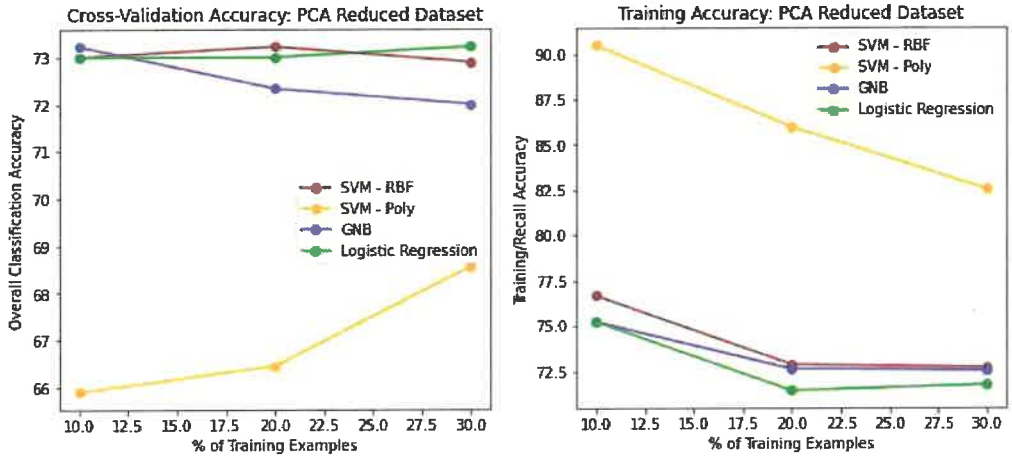

Figure 6. Cross-validation and training accuracy against training size on PCA-reduced dataset.

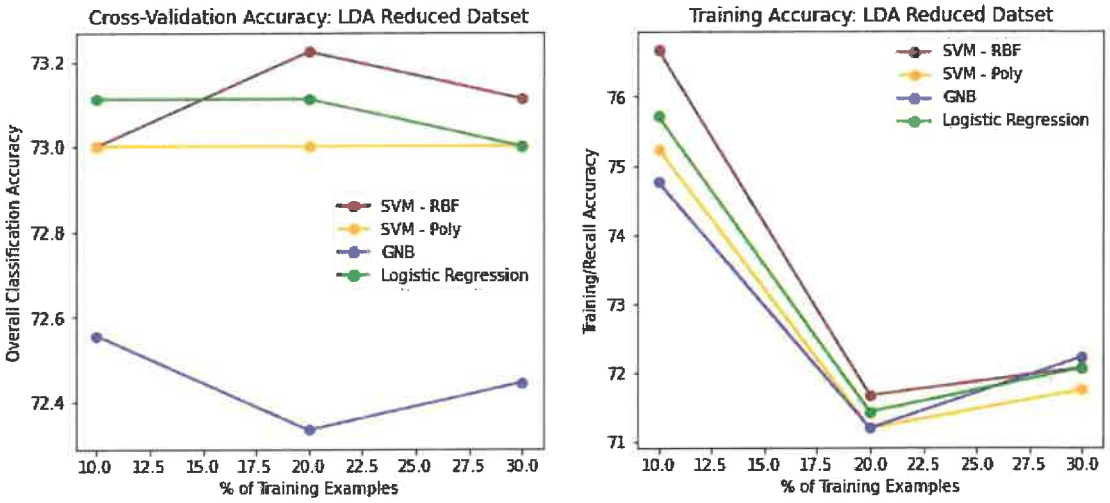

Figure 7. Cross-validation and training accuracy against training size on LDA-reduced dataset.

Confusion Matrix Analysis

In order to evaluate and compare the strength of each model, confusion matrices were generated for each learning technique on the original, PCA-reduced, and LDA-reduced datasets for both 10% and 30% training (**Figures A.1-6**). Three performance metrics were then calculated from the confusion matrices:

1. Accuracy: the ability of the model to correctly classify both binding and nonbinding conformations
2. Sensitivity: the ability of the model to correctly classify binding conformations
3. Selectivity: the ability of the model to correctly classify nonbinding conformations

The results are tabulated below for the original dataset (**Table 1 and 2**), the PCA-reduced dataset (**Table 3 and 4**) and the LDA-reduced datasets (**Table 5 and 6**).

Table 1. Confusion matrix analysis of supervised learning techniques on original ADORA2A dataset with 10% training size.

	SVM - RBF	SVM - Poly	Gaussian Naive Bayes	Logistic Regression
Accuracy	0.713	0.614	0.710	0.715
Sensitivity	0.001	0.389	0.541	0.069
Selectivity	0.996	0.704	0.660	0.968

Table 2. Confusion matrix analysis of supervised learning techniques on original ADORA2A dataset with 30% training size.

	SVM - RBF	SVM - Poly	Gaussian Naive Bayes	Logistic Regression
Accuracy	0.718	0.653	0.610	0.710
Sensitivity	0.007	0.307	0.587	0.125
Selectivity	0.997	0.789	0.619	0.941

Table 3. Confusion matrix analysis of supervised learning techniques on PCA-reduced ADORA2A dataset with 10% training size.

	SVM - RBF	SVM - Poly	Gaussian Naive Bayes	Logistic Regression
Accuracy	0.717	0.631	0.699	0.715
Sensitivity	0.007	0.245	0.154	0.008
Selectivity	0.999	0.784	0.916	0.995

Table 4. Confusion matrix analysis of supervised learning techniques on PCA-reduced ADORA2A dataset with 30% training size.

	SVM - RBF	SVM - Poly	Gaussian Naive Bayes	Logistic Regression
Accuracy	0.717	0.695	0.710	0.715
Sensitivity	0.007	0.108	0.147	0.044
Selectivity	0.997	0.926	0.932	0.979

Table 5. Confusion matrix analysis of supervised learning techniques on LDA-reduced ADORA2A dataset with 10% training size.

	SVM - RBF	SVM - Poly	Gaussian Naive Bayes	Logistic Regression
Accuracy	0.718	0.716	0.717	0.718
Sensitivity	0.025	0.000	0.042	0.021
Selectivity	0.993	1.000	0.985	0.994

Table 6. Confusion matrix analysis of supervised learning techniques on LDA-reduced ADORA2A dataset with 30% training size.

	SVM - RBF	SVM - Poly	Gaussian Naive Bayes	Logistic Regression
Accuracy	0.719	0.718	0.709	0.713
Sensitivity	0.061	0.000	0.221	0.140
Selectivity	0.978	1.000	0.901	0.939

On first glance, when simply looking at overall accuracy, the LDA-reduced dataset seems to perform slightly better than the PCA-reduced dataset, which performs slightly better than the original dataset. In addition, SVM-RBF seems to be the best overall supervised learning technique across most of the models. This evaluation remains the same when considering selectivity, with the dimensionality-reduced datasets performing much better than the original dataset with the SVM-Poly and GNB techniques in this respect.

When considering sensitivity, though, the evaluation is much different. In this respect the original dataset far outperformed the PCA-reduced dataset, which slightly outperformed the LDA-reduced dataset. Both dimensionality-reduction techniques had extremely low and even zero sensitivities across the methods, indicating that the overall accuracy was majorly or entirely due to the classification of nonbinding conformations. This presents a serious issue as the goal of this model development is to identify potential drug candidates, so in a trade-off between selectivity and sensitivity the latter is greatly preferred. Given this insight a new recommendation for the best performing model would be the Gaussian Naive Bayes on the original dataset with

30% training size. While this model has the least overall accuracy, it has significantly higher sensitivity which is more important for this specific application.

Conclusion

This project attempted to develop a machine learning model for classifying protein conformations as binding or non-binding. Various dimensionality reduction and supervised learning techniques were performed on the ADORA2A protein conformation dataset, after which the models were evaluated using performance metrics generated from the confusion matrices. After evaluation there appeared to be a tradeoff between overall accuracy and sensitivity, likely due to the class imbalance of the dataset which contains significantly more non-binding conformations as opposed to binding conformations. Given that the goal of this model is to detect potential drug candidates, sensitivity to binding was determined to be the more important metric and therefore the recommended model is the Gaussian Naive Bayes classifier on the original dataset with 30% training data. This model has approximately 60% overall accuracy as well as 60% sensitivity. While this is not ideal, further improvement could be made by compensating for the class imbalance and attempting other transformations that would optimize both sensitivity and selectivity. If these changes prove to be effective, there is great potential for developing a machine learning model to classify binding affinities of protein conformations which would be a significant improvement in the efficiency of computational drug discovery.

References

[1] Durrant, J. D., & McCammon, J. A. (2011). Molecular dynamics simulations and drug discovery. *BMC Biology*, *9*(71). https://doi.org/10.1186/1741-7007-9-71

[2] Akondi, V. S., Menon, V., Baudry, J., & Whittle, J. (2022). Novel Big Data-Driven Machine Learning Models for Drug Discovery Application. *Molecules*, *27*(3). https://doi.org/10.3390/molecules27030594

[3] *ADORA2A adenosine A2a receptor [Homo sapiens (human)]*. (2023, March 23). National Center for Biotechnology Information.

[4] Evangelista Falcon, W., Ellingson, S. R., Smith, J. C., & Baudry, J. (2019). Ensemble Docking in Drug Discovery. How Many Protein Configurations from Molecular Dynamics Simulations are Needed to Reproduce Known Ligand Binding? *Journal of Physical Chemistry B*, *123*(25), 5189–5195. https://doi.org/10.1021/acs.jpcb.8b11491

[5] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access*, *8*, 54776–54788. https://doi.org/10.1109/ACCESS.2020.2980942

[6] Learned-Miller, E. G. (2014). *Introduction to Supervised Learning*.

[7] Karatzoglou, A., Meyer, D., Wien, W., & Hornik, K. (2006). Support Vector Machines in R. *Journal of Statistical Software* , *15*(9). http://www.jstatsoft.org/

[8] Jahromi, A. H., & Taheri, M. (2017). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. *2017 Artificial Intelligence and Signal Processing Conference (AISP)*, 209–212. https://doi.org/10.1109/AISP.2017.8324083

[9] LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395–2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658

Appendix

Table A.1. Protein conformation descriptors calculated for the ADORA2A protein dataset [2].

Property	Description
pro_mass	Protein Mass
pro_pI_3D	Structure-based pI Prediction
pro_coeff_fric	Frictional Coefficient
pro_coeff_diff	Diffusion coefficient
pro_r_gyr	Radius of Gyration
pro_r_solv	Hydrodynamic Radius
pro_sed_const	Sedimentation Constant
pro_eccen	Protein Eccentricity
pro_asa_vdw	Water Accessible Surface Area
pro_asa_hyd	Hydrophobic Surface Area
pro_asa_hph	Hydrophilic Surface Area
pro_volume	Protein Volume
pro_mobility	Protein Mobility
pro_helicity	Protein Helix Ratio
pro_henry	Henry's Function f(ka)
pro_net_charge	Protein Net Charge
pro_app_charge	Protein Charge at Debye Length
pro_dipole_moment	Protein Dipole Moment
pro_hyd_moment	Hydrophobicity moment
pro_zeta	Zeta Potential
pro_zdipole	Zeta Dipole Moment
pro_zquadrupole	Zeta Quadrupole Moment
pro_patch_hyd	Area of hydrophobic protein patch(es)
pro_patch_hyd_1	Area of largest hydrophobic protein patch(es)
pro_patch_hyd_2	Area of 2 largest hydrophobic protein patch(es)

pro_patch_hyd_3	Area of 3 largest hydrophobic protein patch(es)
pro_patch_hyd_4	Area of 4 largest hydrophobic protein patch(es)
pro_patch_hyd_5	Area of 5 largest hydrophobic protein patch(es)
pro_patch_hyd_n	Count of hydrophobic protein patch(es)
pro_patch_ion	Area of ionic protein patch(es)
pro_patch_ion_1	Area of largest ionic protein patch(es)
pro_patch_ion_2	Area of 2 largest ionic protein patch(es)
pro_patch_ion_3	Area of 3 largest ionic protein patch(es)
pro_patch_ion_4	Area of 4 largest ionic protein patch(es)
pro_patch_ion_5	Area of 5 largest ionic protein patch(es)
pro_patch_ion_n	Count of ionic protein patch(es)
pro_patch_neg	Area of negative protein patch(es)
pro_patch_neg_1	Area of largest negative protein patch(es)
pro_patch_neg_2	Area of 2 largest negative protein patch(es)
pro_patch_neg_3	Area of 3 largest negative protein patch(es)
pro_patch_neg_4	Area of 4 largest negative protein patch(es)
pro_patch_neg_5	Area of 5 largest negative protein patch(es)
pro_patch_neg_n	Count of negative protein patch(es)
pro_patch_pos	Area of positive protein patch(es)
pro_patch_pos_1	Area of largest positive protein patch(es)
pro_patch_pos_2	Area of 2 largest positive protein patch(es)
pro_patch_pos_3	Area of 3 largest positive protein patch(es)
pro_patch_pos_4	Area of 4 largest positive protein patch(es)
pro_patch_pos_5	Area of 5 largest positive protein patch(es)
pro_patch_pos_n	Count of positive protein patch(es)

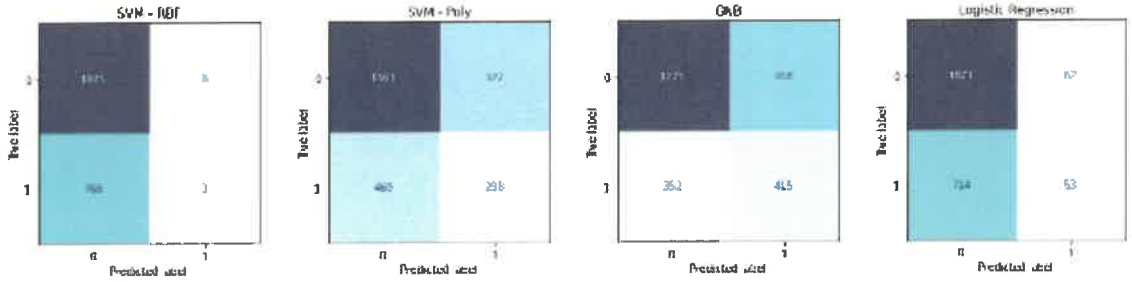

Figure A.1. Confusion matrices for SVM with RBF kernel, SVM with Poly kernel, GNB, and logistic regression on original ADORA2A dataset with 10% training size.

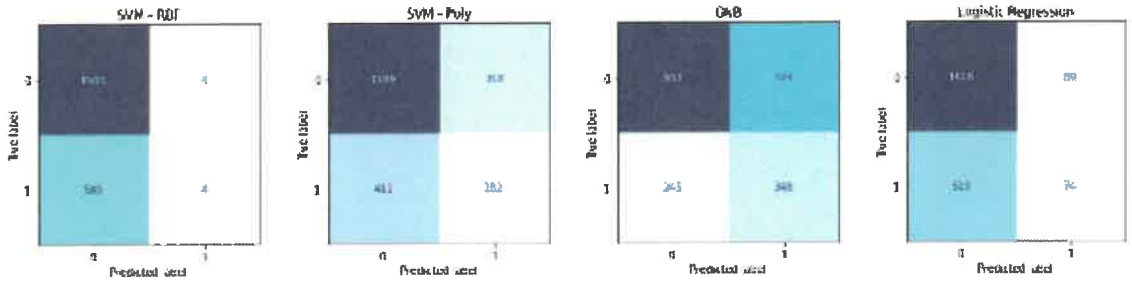

Figure A.2. Confusion matrices for SVM with RBF kernel, SVM with Poly kernel, GNB, and logistic regression on original ADORA2A dataset with 30% training size.

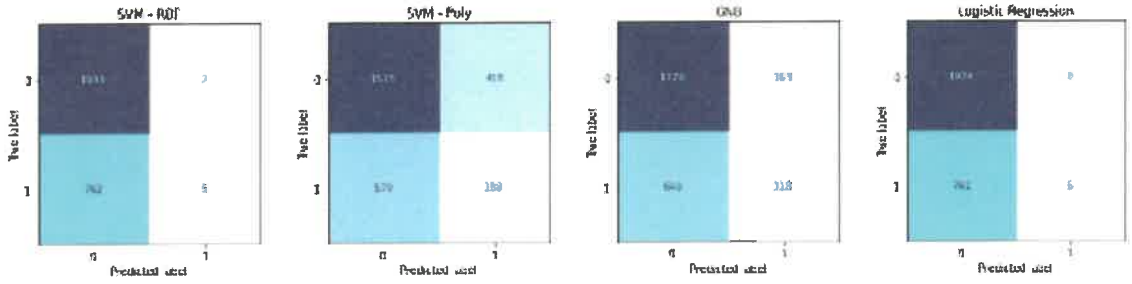

Figure A.3. Confusion matrices for SVM with RBF kernel, SVM with Poly kernel, GNB, and logistic regression on PCA-reduced ADORA2A dataset with 10% training size.

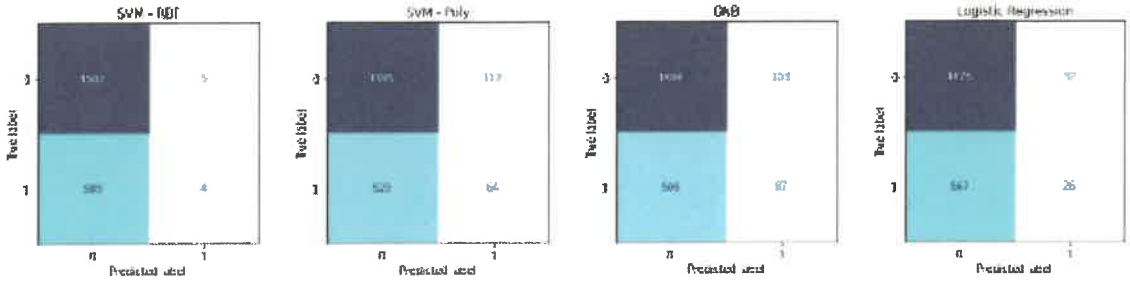

Figure A.4. Confusion matrices for SVM with RBF kernel, SVM with Poly kernel, GNB, and logistic regression on PCA-reduced ADORA2A dataset with 30% training size.

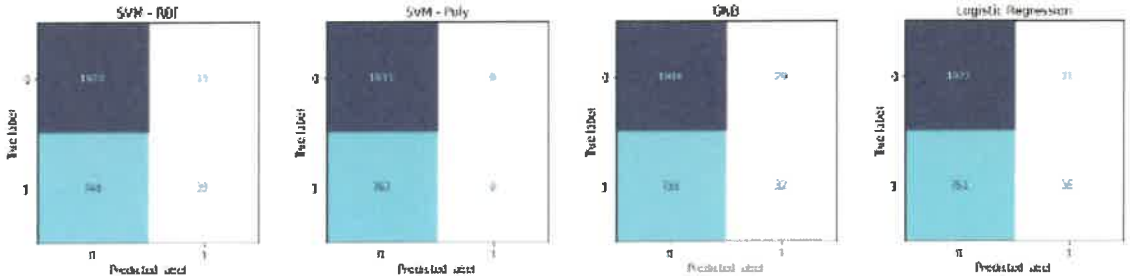

Figure A.5. Confusion matrices for SVM with RBF kernel, SVM with Poly kernel, GNB, and logistic regression on LDA-reduced ADORA2A dataset with 10% training size.

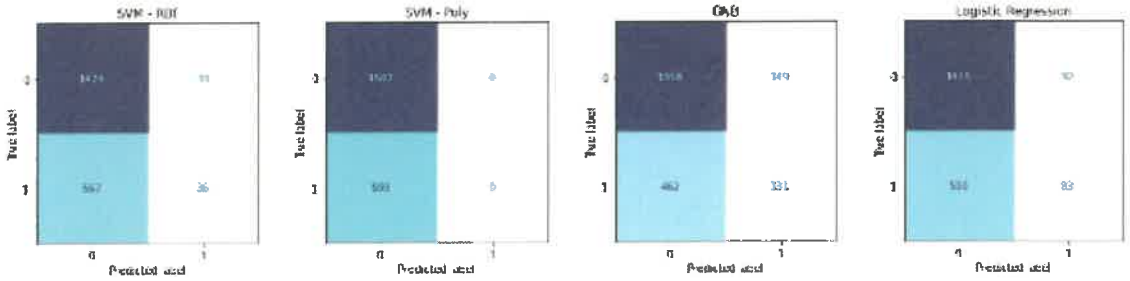

Figure A.6. Confusion matrices for SVM with RBF kernel, SVM with Poly kernel, GNB, and logistic regression on LDA-reduced ADORA2A dataset with 30% training size.

Asani Honors Capstone (2) (3)

Final Audit Report 2023-04-28

Created:	2023-04-28
By:	Eliza Asani (ela0007@uah.edu)
Status:	Signed
Transaction ID:	CBJCHBCAABAA_OkiOF0qeJBzmhISnVAllN4un-5H0pdz

"Asani Honors Capstone (2) (3)" History

Document digitally presigned by Letha Etzkorn (etzkorl@uah.edu)
2023-04-28 - 6:47:26 PM GMT

Document created by Eliza Asani (ela0007@uah.edu)
2023-04-28 - 6:57:30 PM GMT

Document emailed to Eliza Asani (elizal.asani@gmail.com) for signature
2023-04-28 - 6:58:14 PM GMT

Email viewed by Eliza Asani (elizal.asani@gmail.com)
2023-04-28 - 6:58:24 PM GMT

Document e-signed by Eliza Asani (elizal.asani@gmail.com)
Signature Date: 2023-04-28 - 6:58:49 PM GMT - Time Source: server

Agreement completed.
2023-04-28 - 6:58:49 PM GMT