Machine Learning in Breast Cancer Drug Discovery: A Bioinformatics Approach

Vaibhav Ramakrishnan
Applied Modelling and Quantitative Methods
Trent University
Peterborough, ON, Canada
vaibhavramakrishnan@trentu.ca

***Abstract*— Breast cancer - a prevalent and life-threatening disease, demands novel approaches to medication discovery. Current approaches are showing some challenges in terms of sluggishness and inefficiency, making it difficult to identify prospective therapeutic candidates. This study employs a combination of bioinformatics and machine learning to address this critical issue. A quantitative structure-activity relationship (QSAR) model that is specific to breast cancer medication research is created. Unlike earlier studies, this study incorporates data gathering with a focus on the aromatase protein, and user-friendly deployment through a web application. A simple and interpretable model using the Random Forest algorithm is developed in comparison with other machine learning techniques to identify promising medication candidates, advance the field of breast cancer therapy, and inspire new data-driven approaches. The R-squared values for the training and testing sets are 0.75 and 0.21, respectively, which shows that Random Forest is the reasonable and suitable method. Data gathering and preprocessing, exploratory data analysis, machine learning model development, and deployment are all parts of our technique. The combination of robust QSAR models, key insights, and user-friendly tools will together build a bioactivity prediction application that is supposed to significantly enhance the efficacy of breast cancer treatments.**

Keywords*— *breast cancer, aromatase, ChEMBL, Random Forest, bioactivity prediction application

I. INTRODUCTION

Breast cancer has become the world's most prevalent cancer recently with 2.3 million women diagnosed with breast cancer in 2020, resulting in 685.000 global deaths [1]. While breast cancer occurs worldwide and, in both genders, it is more prevalent in women. This is due to the presence of an enzyme called aromatase, which is distributed across various tissues in the human body. Aromatase facilitates the conversion of other hormones into estrogen, which is a primary factor that stimulates breast cancer growth and is generally produced in higher quantities in females compared to males.

Aromatase inhibitors which effectively reduce the production of estrogen are commonly used in the treatment of estrogen receptor-positive breast cancer in postmenopausal women [2]. With the development of technology and machine learning in healthcare, the prediction of aromatase inhibitors has been developed using several techniques in terms of the Efficient Linear Method [3], Bayesian [4], and Support Vector Machine [5]. Although these approaches have shown promise, there are still limitations to their scope. Therefore, there are opportunities for further improvements by developing a simple and user-friendly model that can identify prospective therapeutic candidates effectively.

In this study, we collect bioactivity data from the ChEMBL database and perform an exploratory data analysis based on Lipinski's descriptors to determine the drug-likeness of compounds. After that, we propose an approach using a Random Forest algorithm to predict the aromatase inhibitory activity and compare it with other machine learning methods. Finally, a bioactivity prediction application will be introduced, which can be used to predict the aromatase inhibitory activity with molecule input.

II. PREVIOUS WORK

Before proceeding with our study, we analyzed previous research done in the field of bioinformatics. Various studies have explored the application of machine learning in predicting bioactivity. Most studies involved the usage of diverse datasets from different sources, data preprocessing, exploratory data analysis, and statistical analysis. The study [4], used aromatase protein as a target and used machine learning models to predict its inhibition. Their study also compared the performance of machine learning algorithms to internal five-fold cross-validation statistics of training data. This helps give direction to our research, as we compare Random Forest with different other algorithms to find out the optimal model. Our study, which builds on the work reported in [5], takes a novel approach by addressing a significant gap in the previous research. While [5] predominantly employs a Support Vector Machine (SVM) predictor, it emphasizes protein patterns, amino acid content, and dipeptide compositions. However, our research goes beyond these factors by looking into the relevance of pIC50 values. In contrast to earlier research, we concentrate on the bioinformatics element of representing a substance's ability to block aromatase, providing useful insights into breast cancer therapy and the creation of critical medications in this sector. Bioinformatics is crucial in identifying new drug candidates, predicting drug toxicity, and finding an efficient drug design [6]. In existing research, some studies did not consider pharmacokinetic properties of molecules such as absorption, distribution, metabolism, and excretion (ADME) profiles (see [7]), which our study ensures by the calculation of Lipinski's descriptors which is based on pharmacokinetic properties of compounds [8].

III. METHODOLOGY

Our holistic approach to bioactivity prediction draws inspiration from Chanin Nantasenamat (Data Professor), whose YouTube channel provided invaluable insights (see [9]). The instructional videos by Chanin Nantasenamat played a crucial role in guiding us through coding, model building, and application development, significantly enriching our research endeavors. In this section, we outline our

methodology, encompassing steps such as data gathering from the ChEMBL database, data pre-processing, exploratory data analysis, and computation of Lipinski's descriptors. An important aspect of our process involves converting IC50 to pIC50, and we elucidate its significance. Complementing our approach, we employ the Mann-Whitney U-Test for statistical analysis. These subsections comprehensively address key aspects before progressing to the subsequent stage of machine learning model building.

A. *Dataset Collection*

The target protein of interest to build the machine learning model (QSAR) for this breast cancer research was Aromatase, which was collected from the ChEMBL database. The ChEMBL database contains the biological activity data of millions of different compounds (see [10]). The chembl_webresource_client Python library allows us to download the biological activity data from the ChEMBL database. In this research, Aromatase is the target protein on which the drug or the compound will act on. Biologically, the drugs will come in contact with the protein and induce a modulatory or inhibitory activity towards it. The bioactivity measure chosen for this research was inhibitory concentration (IC50). It represents the concentration of a compound or drug that is required to inhibit the protein of interest by 50%. This aids in the discovery of new drug candidates for breast cancer treatment.

B. *Data Preprocessing*

Before proceeding with the exploratory data analysis and model building, the data preprocessing phase is performed to deal with missing values, duplicate values, and outliers.

- The compounds are labeled as either active, intermediate, or inactive.
- Compounds with values less than 1000 nM (nanomolar) are active, while those greater than 10,000 nM are inactive. The compounds between 1000 nM and 10,000 nM are referred to as intermediate. For example, see [11], where the author uses a scoring function, RDOCK, and sets a threshold of 50 kcal/mol to find out the binding affinity of compounds.
- The resultant data frame consists of the molecule ChEMBL IDs, the canonical SMILES (Simplified Molecular Input Line Entry System) notations, which are sequential in nature [12], and the standard values (or the potency of a drug) of the respective compounds.

C. *Exploratory Data Analysis*

The Exploratory Data Analysis consisted of the preparation of SMILES notation, calculation of Lipinski's descriptors, and analysis using visualizations and statistical tests (Mann-Whitney U Test) to understand the distribution and patterns in our dataset.

D. *Lipinski's Descriptors Calculation*

The term comes from a Pfizer scientist named Christopher Lipinski, who devised a series of guidelines known as the Rule-of-Five or Rule of Thumb [13]. This is used to assess a compound's drug-likeness (or drug-like characteristics). The drug-likeness is determined by essential pharmacokinetic features such as ADME: Absorption, Distribution, Metabolism, and Excretion [8],[14] and [15]. Christopher Lipinski gathered a collection of FDA-approved medications that are often delivered orally [16]. According to his findings, the four descriptors utilized had comparable values in multiples of five, as follows:

- Molecular weight < 500 Dalton
- Octanol-water partition coefficient (LogP) < 5
- Hydrogen bond donors < 5
- Hydrogen bond acceptors < 10

The Simplified Molecular Input Line Entry System (SMILES) notation is a standardized chemical format representing the chemical structure or information of molecules. The Lipinski descriptors calculated using Python Library functions, use the SMILES notation as input.

E. *Convert IC50 to pIC50*

The IC50 values have an uneven distribution of values in our dataset. To make the distribution more even, a negative logarithmic base 10 transformation is done to convert to pIC50 values. This is a common practice in bioinformatic studies [17], which serves the purpose of making the dataset more amenable to statistical analysis and modeling. The pIC50 values are subsequently used as the target variables for model building and analysis.

For a simple comparison between the bioactivity classes i.e., active, and inactive compounds, the intermediate class is deleted from the curated dataset. Using Python libraries, we can use visualizations to compare the distribution of active and inactive molecules (see Fig. 1).

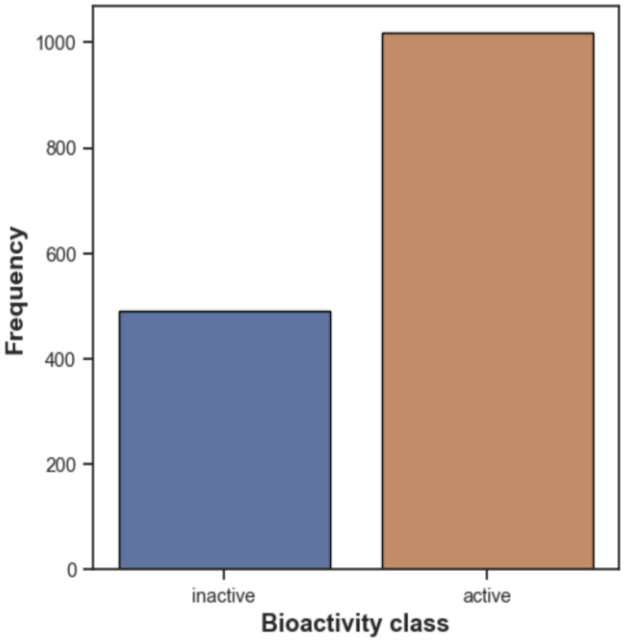

Fig. 1. Bioactivity class distribution (Active v/s Inactive compounds)

The 2 bioactivity classes span similar chemical spaces as evidenced by the scatterplot of Molecular weight v/s LogP or the solubility of a compound, (see Fig. 2). The seaborn Python library helps create these visualizations.

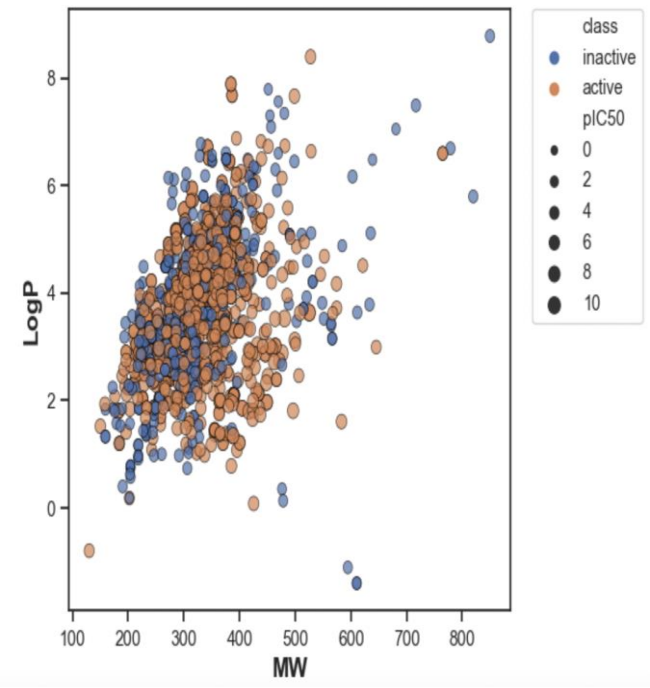

Fig. 2. Molecular weight (MW) v/s LogP (solubility)

To further understand the distribution of the inactive and the active classes, we use a boxplot (see Fig. 3). The threshold values of 5 nM and 6 nM for the pIC50 value help us distinguish between the 2 classes. With pIC50 values, more than 6 nM are termed as active, and less than 5 nM are termed as inactive.

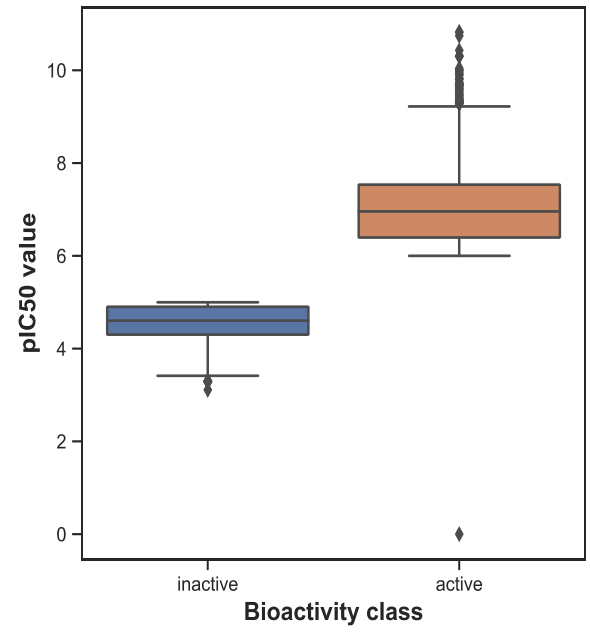

Fig. 3. Boxplot distribution of active and inactive class

F. Mann-Whitney U Test

As part of our statistical analysis, we used the Mann-Whitney U Test to test the statistical significance between the difference of two distributions of active and inactive classes (see Table I).

TABLE I. MANN-WHITNEY U TEST RESULTS FOR ACTIVE V/S INACTIVE CLASS

Descriptor	Statistics	p	alpha	Interpretation
pIC50	500856.0	1.632980e-217	0.05	Different distribution (reject H0)

As per the results shown for the statistical analysis in Table I, the p-value is lower than the significance level alpha (which is 0.05), meaning that we can reject the null hypothesis. This helps us conclude that active and inactive compounds have different distributions.

To further understand the distribution of 4 Lipinski descriptors, we use visualizations and statistical analysis as well.

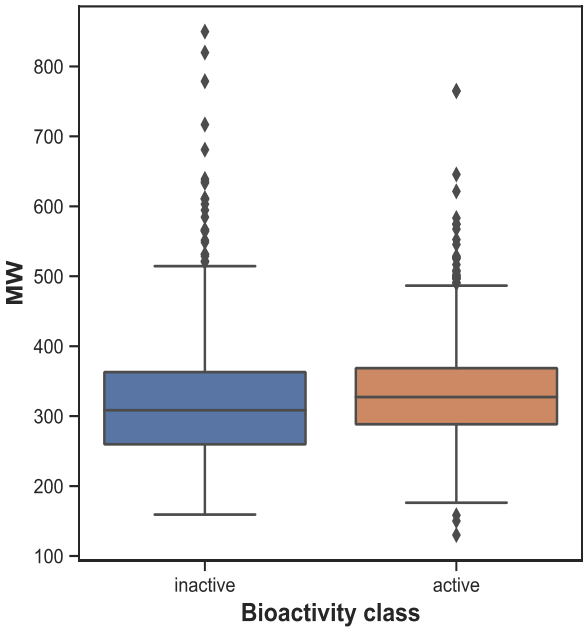

Fig. 4. Boxplot of active v/s inactive class for Molecular Weight (MW)

TABLE II. MANN-WHITNEY U TEST RESULTS FOR ACTIVE V/S INACTIVE CLASS FOR MOLECULAR WEIGHT (MW)

Descriptor	Statistics	p	alpha	Interpretation
MW	284791.0	0.000018	0.05	Different distribution (reject H0)

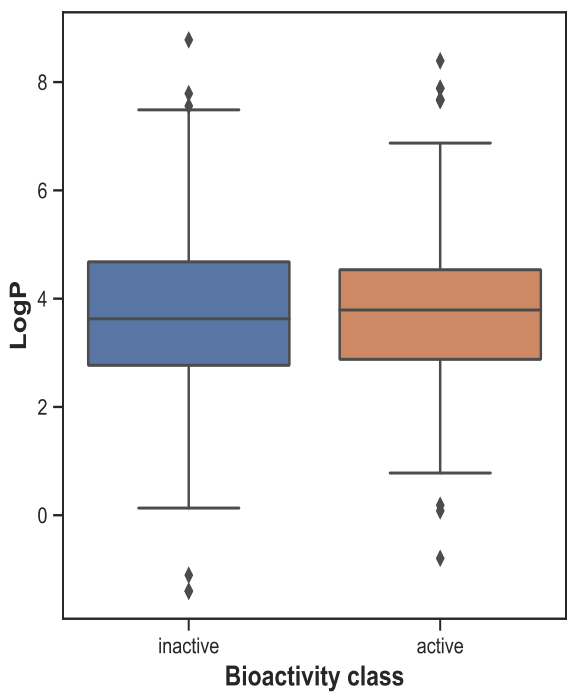

Fig. 5. Boxplot of active v/s inactive class for Solubility (LogP)

TABLE III. MANN-WHITNEY U TEST RESULTS FOR ACTIVE V/S INACTIVE CLASS FOR SOLUBILITY (LOGP)

Descriptor	Statistics	p	alpha	Interpretation
LogP	252241.5	0.843704	0.05	Same distribution (fail to reject H0)

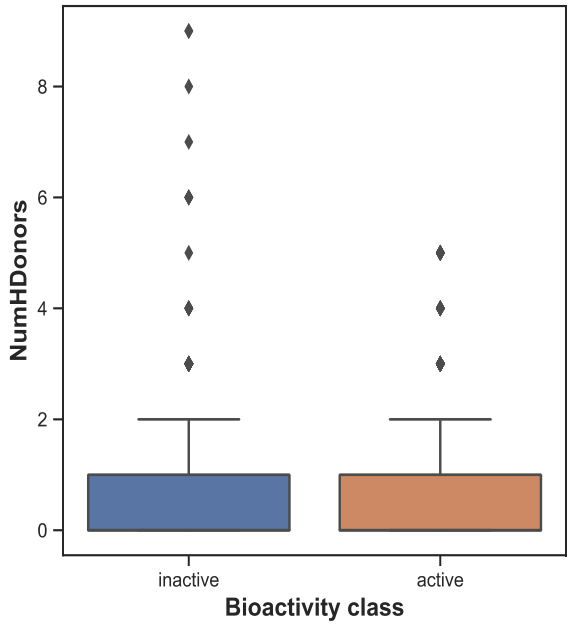

Fig. 6. Boxplot of active v/s inactive class for NumHDonors

TABLE IV. MANN-WHITNEY U TEST RESULTS FOR ACTIVE V/S INACTIVE CLASS FOR NUMHDONORS

Descriptor	Statistics	p	alpha	Interpretation
NumHDonors	234631.0	0.023067	0.05	Different distribution (reject H0)

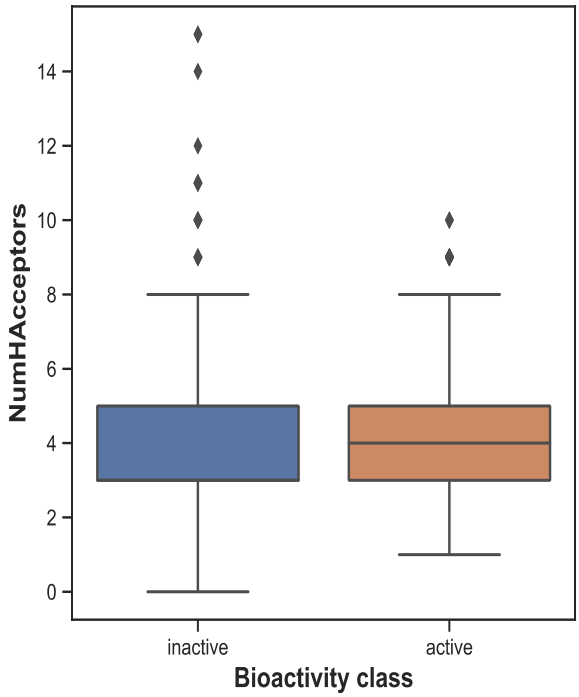

Fig. 7. Boxplot of active v/s inactive class for NumHAcceptors

TABLE V. MANN-WHITNEY U TEST RESULTS FOR ACTIVE V/S INACTIVE CLASS FOR NUMHACCEPTORS

Descriptor	Statistics	p	alpha	Interpretation
NumHAcceptors	270837.0	0.009792	0.05	Different distribution (reject H0)

As per the results from Mann-Whitney U Test, we are able to conclude that 3 of the 4 Lipinski's descriptors exhibited statistically significant differences between the active and inactive compounds except for LogP (or solubility) as seen in Table III, which shows the same distribution between the 2 classes i.e. active and inactive., while the other 3 descriptors (MW, NumHDonors, NumHAcceptors) show statistically significant difference between actives and inactive. While this violation of Rule-of-5 may be a slight deviation from an ideal drug candidate, a study in [13] discusses how nearly 18% of drugs deviate from Rule-of-5 with one violation, while few others violate 2 or 3 descriptors. But, it does not necessarily mean that the compound cannot be developed as a drug.

IV. MACHINE LEARNING MODEL BUILDING

Before starting the model-building phase, we prepare the dataset. This involved calculating the molecular descriptors that are essentially quantitative descriptions of the compounds. These describe the local features of the molecules, essentially representing the unique building blocks that constitute each molecule. We focus on the canonical SMILES notations and molecule ChEMBL IDs for this preparation. The PaDEL descriptor software file helps clean the chemical structure by removing all the salt and organic acids so that there are no impurities. Ultimately, the PaDEL descriptor helps compute the molecular fingerprints [18]. The final dataset contains the fingerprints and their respective pIC50 values.

A. *Random Forest Model*

There were 881 input fingerprints in our dataset. Each molecule has a unique fingerprint (like humans). This enables the machine learning system to learn from the compound's unique molecular features, as well as helps create a model that will be able to distinguish between compounds that are active v/s inactive. This is the goal of our model-building phase, as we look to identify which functional group or fingerprint is essential for the design of a potent or effective drug against breast cancer. The target variable for the prediction is pIC50 [17], [19].

Using a data split of 80/20, we successfully built a Random Forest Regression Model. The model achieved an R-squared score of close to 38% which indicates that the model is able to account for a substantial portion of variability in the test dataset.

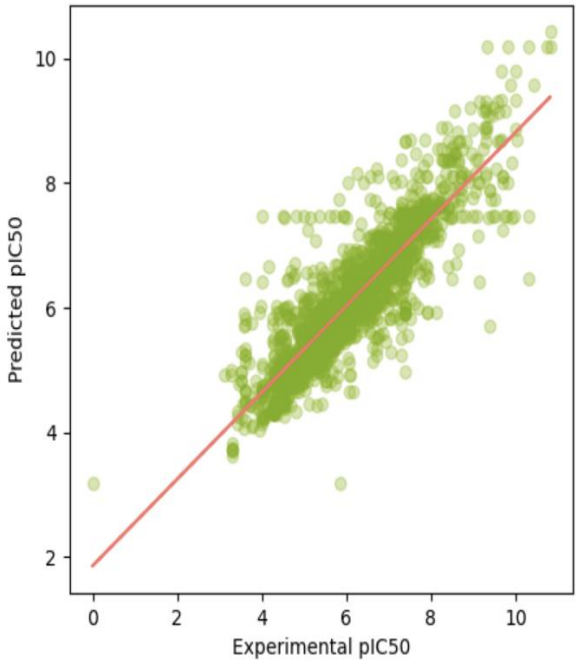

Fig. 8. Scatterplot of Predicted v/s Experimental pIC50 values for the Random Forest Model

The scatterplot created by our Random forest model, as shown in Fig. 8, reveals that the cluster of points along the line of best fit indicates that our model had a good fit to the training data, allowing us to infer that Random forest is an excellent model for further analysis and application development.

B. *Model comparison*

The Lazypredict Python library is used to compare several machine-learning algorithms [20]. By performing a data split of 80/20 for the training and testing data, with the help of sci-kit-learn libraries, we are able to compare the performance of close to 40 machine-learning algorithms. See Fig. 9 & Fig. 10 for a performance comparison of test and training sets.

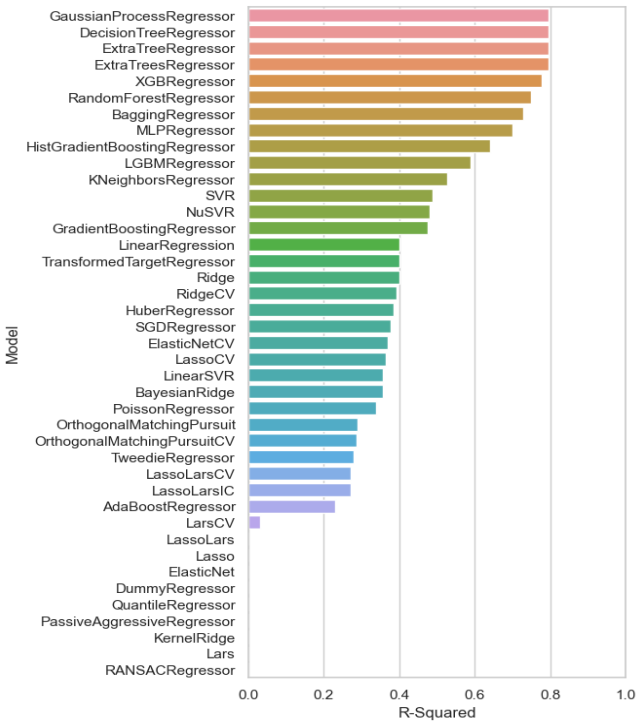

Fig. 9. Bar Plot for R-Squared values for training set

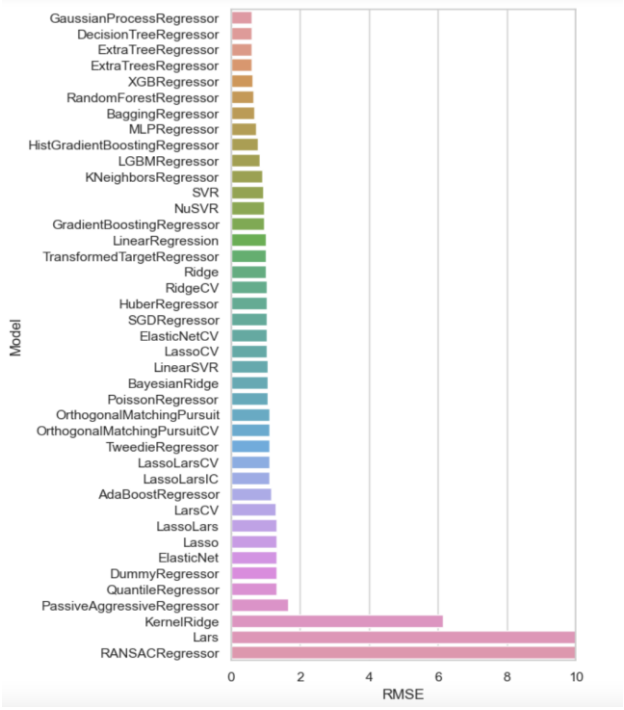

Fig. 10. Bar Plot for RMSE values for training set

As per Fig. 9, we can conclude that the R-Square value of GaussianProcessRegressor is highest, while the Random Forest is close with a value of close to 0.75 for the training set. This proves that these models are able to explain a high percentage of variance in the target variable and capture underlying patterns in the data. The results from Fig. 10 show the RMSE values of GaussianProcessRegressor, DecisionTreeRegressor, XGBRegressor, and Random Forest are very close to each other and fall in the range of 0 to 2. This suggests that these models have low errors in predicting the target variable, indicating good accuracy and precision [19].

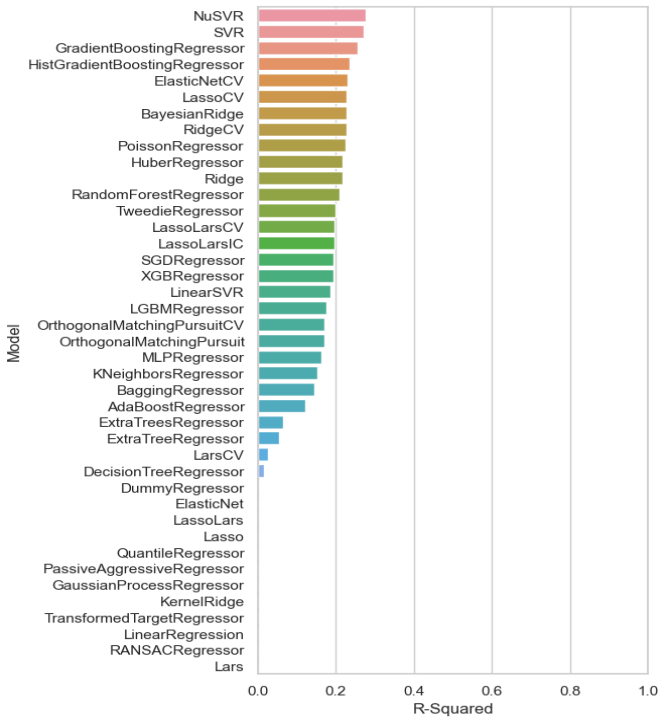

Fig. 11. Bar Plot for R-Squared values for test set

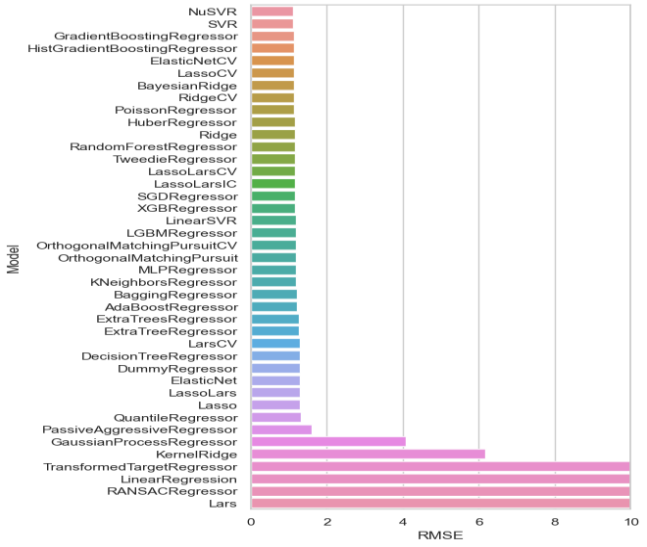

Fig. 12. Bar Plot for RMSE values for test set

The testing set results shown in Fig. 11 & Fig. 12, the RMSE values for NuSVR, SVR, GradientBoostingRegressor, and RandomForestRegressor are all very close to each other and fall in the range of 0 to 2. This suggests that these models have low errors in predicting the target variable on the testing set, indicating good accuracy and precision. The R-squared values for NuSVR and SVR are the highest, approximately around 0.25. This indicates that these models explain a higher percentage of the variance in the target variable on the testing set compared to the other models. However, an R-squared value of 0.25 suggests that there is still room for improvement in explaining the variance. The R-squared value for RandomForestRegressor is around 0.21. While not as high as NuSVR and SVR, it still suggests that the model provides a reasonable fit to the testing data.

C. Optimal Model

A few models like NuSVR, SVR, DecisionTreeRegressor, etc. performed slightly better for RMSE and R-squared values than the Random Forest model. These differences however are not significant, and we choose to proceed with the Random Forest model as our most optimal model. Random Forest shows a low RMSE value on the training set, which proves that it performed well in fitting the training data. Besides that, it had a reasonable RMSE on the testing set, which means that Random Forest generalizes well to unseen data. The Random Forest model had a decent R-squared value for the test set around 0.21. While this value is not the highest, it is still a reasonable fit for the testing data. The Random Forest can handle different data types which includes categorical and numerical, as well as it is also less prone to overfitting [21]. Furthermore, it can perform well for large datasets, which helps us conclude that it is the optimal choice.

V. APPLICATION ARCHITECTURE

The Random Forest machine-learning model built on the Aromatase bioactivity dataset is converted into a web application. This helps allow users to make predictions on our machine-learning model for the target protein of their interest (see Fig. 13).

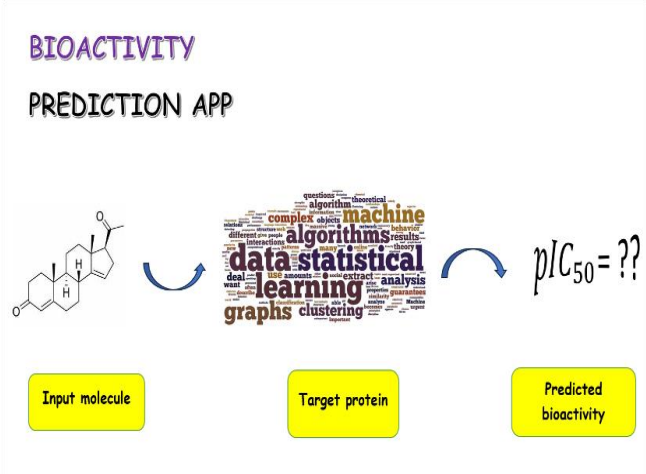

Fig. 13. Bioactivity Prediction Application Homepage

The application calculates the molecular descriptors of the input data which contains the SMILES notation of compounds along with their respective ChEMBL IDs (see Fig. 14).

example_aromatase.txt

```
CCn1c(C(c2ccc(F)cc2)n2ccnc2)c(C)c2cc(Br)ccc21 CHEMBL431859
CCn1cc(C(c2ccc(F)cc2)n2ccnc2)c2ccccc21 CHEMBL113637
Clc1ccccc1Cn1cc(Cn2ccnc2)c2ccccc21 CHEMBL112021
CCn1ccc2cc(C(c3ccc(F)cc3)n3ccnc3)ccc21 CHEMBL41761
Cn1cc(C(c2ccc(F)cc2)n2ccnc2)c2cc(Br)ccc21 CHEMBL111868
```

Fig. 14. Contents of the input file with SMILES notation

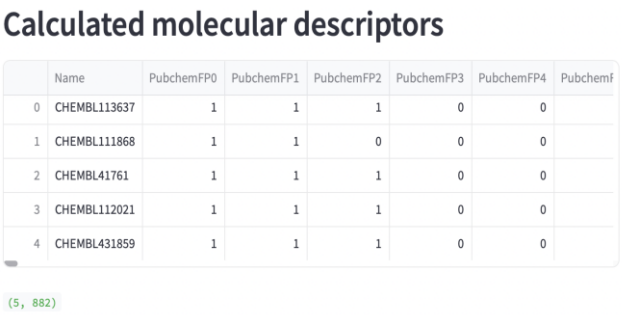

Fig. 15. Molecular descriptors calculated using the Bioactivity prediction application

The output in Fig. 15 for the example aromatase input file shows there are 5 input molecules and 881 molecular fingerprints which are the PubChem fingerprints. Using our machine learning model, we reduce the number of descriptors from 881 to 248 (see Fig. 16). This is done to remove the low variance or excessive redundant features from the dataset, to allow us to build the model much quicker.

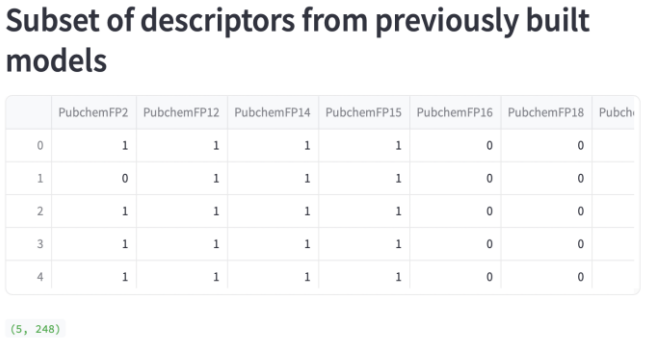

Fig. 16. Subset of descriptors using previously built model

These 248 descriptors, together with their accompanying ChEMBL IDs, are utilized as an X variable in the Python function to forecast the pIC50 values in our prediction output. The tool also allows users to download their predictions (see Fig. 17).

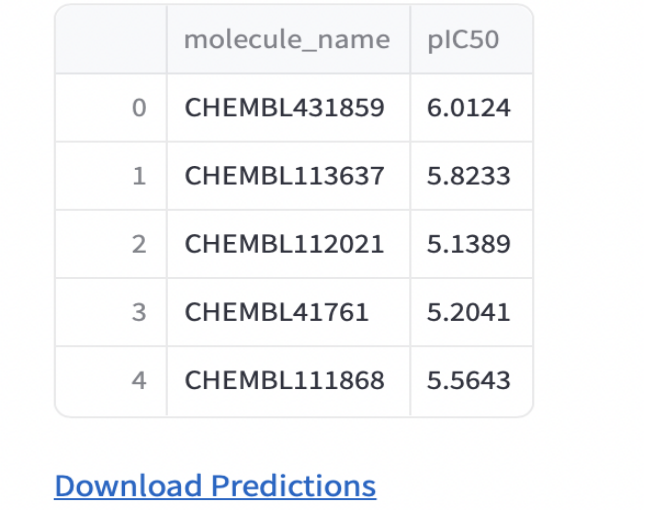

Fig. 17. Prediction Output from the Bioactivity Prediction Application

VI. RESULTS

The Random Forest model attained an R-squared value of 0.7652, indicating the proportion of variance in our target variable, pIC50, explained by our input molecular features is 76.52%. Further, the model achieved an MSE score of 0.40 which is reasonable. In addition, when we compared Random Forest with several other machine-learning algorithms, Random Forest did not perform the best for respective training and testing sets, it showed reasonable RMSE and R-squared values. These values, along with the scatterplot results (see Fig. 8) prove that Random Forest was an effective model in predicting bioactivity against aromatase. The model-building part was complex due to the complexity of the dataset, we needed to understand the SMILES notation and created our model accordingly. Finally, we successfully used this Random Forest model to build our Bioactivity Prediction Application. This user-friendly application helps researchers and users upload their own bioactivity data to predict inhibition or the pIC50 value (see Fig. 13). The application is our contribution towards breast cancer drug discovery and specifically targeting aromatase inhibition.

VII. CONCLUSION

In this study, we obtained our target protein dataset for Aromatase from the ChEMBL database. Using this dataset, we built a Random Forest model and compared it against 41 machine learning techniques. Random Forest was the optimal model based on the major metrics of R-squared and RMSE values for the training and testing sets, as well as additional variables such as our model's capacity to handle numerical and categorical data characteristics, overfitting, and handling of big datasets. Although several models outperformed Random Forest in terms of R-squared and RMSE, the differences were not statistically significant. Finally, we utilized this model to create our online application – Bioactivity Prediction Application, which was aimed at finding medications that target breast cancer cells. As part of future work, we plan to integrate predictive toxicology and real-time data updates (the most recent compound information) into our web application.

REFERENCES

[1] World Health Organization (WHO), "Breast cancer", 2023, [Online], Available: http://www.who.int/news-room/fact-sheets/detail/breast-cancer

[2] L. Eldridge, MD, "Aromatase Inhibitors to Prevent Breast Cancer Recurrence", Verywellhealth, Jan. 2023, [Online], Available: http://www.verywellhealth.com/aromatase-inhibitors-for-preventing-breast-cancer-recurrence-4153970

[3] W. Shoombuatong, V. Prachayasittikul, V. Prachayasittikul, and C. Nantasenamat, "Prediction of aromatase inhibitory activity using the efficient linear method (ELM)", EXCLI Journal, vol. 14, pp. 452-464, Mar. 2015, doi: http://dx.doi.org/10.17179/excli2015-140

[4] K. M. Zorn et al., "Comparing Machine Learning Models for Aromatase (P450 19A1)," Environmental Science & Technology, vol. 54, no. 23, pp. 15546–15555, Nov. 2020, doi: https://doi.org/10.1021/acs.est.0c05771.

[5] Muthu Krishnan Selvaraj and J. Kaur, "Computational method for aromatase-related proteins using machine learning approach," PLOS ONE, vol. 18, no. 3, pp. e0283567–e0283567, Mar. 2023, doi: https://doi.org/10.1371/journal.pone.0283567.

[6] S. K. Wooller, G. Benstead-Hume, X. Chen, Y. Ali, and F. M. G. Pearl, “Bioinformatics in translational drug discovery,” Bioscience reports, vol. 37, no. 4, 2017, doi: 10.1042/BSR20160180

[7] S. He et al., “Machine Learning Enables Accurate and Rapid Prediction of Active Molecules Against Breast Cancer Cells,” Frontiers in Pharmacology, vol. 12, Dec. 2021, doi: https://doi.org/10.3389/fphar.2021.796534.

[8] H. Mishra, N. Singh, T. Lahiri, and K. Misra, “A comparative study on the molecular descriptors for predicting drug-likeness of small molecules,” Bioinformation, vol. 3, no. 9, pp. 384–388, May 2009, doi: https://doi.org/10.6026/97320630003384.

[9] [freeCodeCamp.org], "Python for Bioinformatics - Drug Discovery Using Machine Learning and Data Analysis," YouTube, Jun. 2021, [Online], Available: https://www.youtube.com/watch?v=jBlTQjcKuaY&t=260s

[10] D. Mendez et al., “ChEMBL: towards direct deposition of bioassay data,” Nucleic Acids Research, vol. 47, no. D1, pp. D930–D940, Nov. 2018, doi: https://doi.org/10.1093/nar/gky1075.

[11] M. Tsuji, “Potential anti‐SARS‐CoV‐2 drug candidates identified through virtual screening of the ChEMBL database for compounds that target the main coronavirus protease,” FEBS Open Bio, May 2020, doi: https://doi.org/10.1002/2211-5463.12875.

[12] J. Arús-Pous et al., “SMILES-based deep generative scaffold decorator for de-novo drug design,” Journal of Cheminformatics, vol. 12, no. 1, May 2020, doi: https://doi.org/10.1186/s13321-020-00441-8.

[13] T. K. Karami, S. Hailu, S. Feng, R. Graham, and H. J. Gukasyan, “Eyes on Lipinski’s Rule of Five: A New ‘Rule of Thumb’ for Physicochemical Design Space of Ophthalmic Drugs,” Journal of Ocular Pharmacology and Therapeutics, vol. 38, no. 1, pp. 43–55, Jan. 2022, doi: https://doi.org/10.1089/jop.2021.0069.

[14] L. Z. Benet, C. M. Hosey, O. Ursu, and T. I. Oprea, “BDDCS, the Rule of 5 and drugability,” Advanced Drug Delivery Reviews, vol. 101, pp. 89–98, Jun. 2016, doi: https://doi.org/10.1016/j.addr.2016.05.007.

[15] A. Mullard, “Re-assessing the rule of 5, two decades on,” Nature Reviews Drug Discovery, vol. 17, no. 11, pp. 777–777, Oct. 2018, doi: https://doi.org/10.1038/nrd.2018.197.

[16] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, “Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings,” Advanced Drug Delivery Reviews, vol. 46, no. 1–3, pp. 3–26, Mar. 2001, doi: https://doi.org/10.1016/s0169-409x(00)00129-0.

[17] José Guadalupe Rosas-Jiménez, M. A. García‐Revilla, A. Madariaga-Mazón, and K. Martínez‐Mayorga, “Predictive Global Models of Cruzain Inhibitors with Large Chemical Coverage,” ACS omega, vol. 6, no. 10, pp. 6722 – 6735, Mar. 2021, doi: https://doi.org/10.1021/acsomega.0c05645.

[18] C. W. Yap, “PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints,” Journal of Computational Chemistry, vol. 32, no. 7, pp. 1466–1474, Dec. 2010, doi: https://doi.org/10.1002/jcc.21707.

[19] I. Aqeel, M. Bilal, A. Majid, and T. Majid, “Hybrid Approach to Identifying Druglikeness Leading Compounds against COVID-19 3CL Protease,” Pharmaceuticals, vol. 15, no. 11, p. 1333, Oct. 2022, doi: https://doi.org/10.3390/ph15111333.

[20] A. E. Eldin Rashed, A. M. Elmorsy, and A. E. Mansour Atwa, “Comparative evaluation of automated machine learning techniques for breast cancer diagnosis,” Biomedical Signal Processing and Control, vol. 86, p. 105016, Sep. 2023, doi: https://doi.org/10.1016/j.bspc.2023.105016.

[21] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, “A comparison of random forest variable selection methods for classification prediction modeling,” Expert Systems with Applications, vol. 134, pp. 93–101, Nov. 2019, doi: https://doi.org/10.1016/j.eswa.2019.05.028.