RESEARCH PAPER

Predictive Modelling of EGFR Binding Affinity : An optimized data driven approach.

Parimala Palli [1], Satyasis Mishra [2], P.Srinivasa Rao [3]

[1]*Research Scholar, Dept. of CSE Centurion University of Technology and Management Bhubaneswar, Odisha, India parimala.rachem@gmail.com*
[2]*Dept. of ECE Centurion University of Technology and Management Bhubaneswar, Odisha, India s.mishra@cutm.ac.in*
[3]*Dept. of CSE Maharaj Vijayaram Gajapathiraj College of Engineering (A) Andhra Pradesh, India psr.sri@gmail.com*

ABSTRACT
This paper, is to facilitate the discovery of Epidermal Growth Factor Receptor (EGFR) inhibitors, which plays a vital role in cancer therapy, cheminformatics and machine learning (ML) are used. Molecular descriptors, such as Lipinski Rule of Five and other structural properties, were estimated using bioactivity data of the ChEMBL database to determine drug-likeness. The predictive performance of the ML models (Random Forest (RF), Gradient Boosting Regressor, and Neural Networks) trained using these descriptors was to predict the EGFR binding affinity with high accuracy. To determine the most successful algorithms, such evaluation metrics as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R 2) were used. Researchers have proven that there is a strong, data-based pipeline of drug-like EGFR inhibitor prioritization, which illustrates the promise of computational tools in facilitating cancer drug discovery, reducing budgets and improving the development of therapeutic agents..
Keywords: Cheminformatics, Machine Learning (ML), Epidermal Growth Factor Receptor (EGFR), Drug Discovery, Bioactivity Data, Molecular Descriptors, Binding Affinity..
How to cite this article: Parimala Palli, Satyasis Mishra, P. Srinivasa Rao. Predictive Modelling of EGFR Binding Affinity: An Optimized Data Driven Approach. Int J Drug Deliv Technol. 2026;16(37s): 682-692. DOI: 10.25258/ijddt.16.37s.88
Source of support: Nil.
Conflict of interest: The author declares no conflict of interest, and this work represents independent academic research conducted in a personal capacity, not associated with any employer or commercial entity.

INTRODUCTION

This article integrates cheminformatics and ML in search of EGFR inhibitors and finds them with high precision. The bioactivity data of ChEMBL database was studied to calculate molecular descriptors which used to estimate the drug-likeness and binding affinity. The workflow fitted regression models including the Random Forest and the Gradient Boosting in order to model bioactivity that ensures the best compounds are selected. To outperform a data-driven pipeline that will uncover EGFR inhibitors, this paper carries out studies with a huge amount of preprocessing, desktop engineering of descriptors, and model evaluation through the tools of Mean Absolute Error (MAE) and R-squared (R^2). In the approach, it has been made clear that computational tools can reduce the cost of drug discovery and expedite the development of effective drug treatment of cancer. This synergistic match between cheminformatics and ML offers a visionary surrogate of addressing the obstacles of drug resistance and specificity of mutation with EGFR-targeted therapies.

1.1 Background on EGFR and Its Role in Cancer

Epidermal Growth Factor Receptor (EGFR) is a transmembrane protein, where the ligand binding of EGFR triggers signaling pathways that control the cellular processes such as proliferation, differentiation, and survival [1]. Overexpression or mutations of the EGFR has been linked to diverse forms of cancer, such as non-small cell lung cancer, colorectal cancer and glioblastoma which result in unregulated increase in cell numbers, and neoplastic progression [2]. Consequently, EGFR is an established oncolytic target to be used in designing targeted therapies [3]. EGFR-targeting small-molecule EGFR inhibitors and monoclonal antibodies have been effective in terms of EGFR inhibition, which has led to better clinical outcomes in patients with EGFR-driven cancers [4].

1.2 Role of Cheminformatics in EGFR Inhibitor Discovery

In this work, we apply the cheminformatics tools to find new possible EGFR inhibitors through the analysis of ChEMBL database chemical data[5]. ChEMBL is a large warehouse of bioactive molecules containing information about the molecular structures, bioactivity and interaction of the molecules with biological targets[6]. A set of molecules with an inhibitory potential would be obtained by asking ChEMBL to find the compounds with a known activity against EGFR, forming the basis of machine learning model training. We determine the drug-likeness of each compound using molecular descriptors, including those of Lipinski Rule of Five (molecular weight, hydrogen bond donors and acceptors, and logo) [7]. This is to enhance the good pharmacokinetic profiles of the chosen candidates which play vital roles in the progression of compounds made in silico to experimental validation and possible application in a clinical setting.

1.3 Machine Learning in Cheminformatics and Drug Discovery

**Author for Correspondence: Parimala Palli*

Machine learning (ML) has been integrated into cheminformatics, which allows the development of predictive models that can be used to correlate chemical structure with biological activity. Regression algorithms and some other ML models can be used to predict the potency of compounds against EGFR using known data as training[8]. In this paper, a comparison of a variety of ML regressor models such as the Random Forest, Gradient Boosting, and neural networks are used to predict the inhibitory potency of compounds. We will use these models to rank compounds that are most likely to be useful in inhibiting the activity of EGFR[9].

Features engineering plays a vital role in cheminformatics because molecular descriptors can represent chemical and structural features that are necessary to biological activity. In this research, the input features will be the molecular descriptors produced by such software as PaDEL, whereas the target output variable will be the pIC50 values, which are to be regressed. The performance of every ML model is measured with the help of such important indicators as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the value of R-squared (R^2) to identify predictive accuracy and reliability.

1.4 Data Preprocessing and Feature Selection for Model Optimization

One of the key steps in the creation of effective ML models is preprocessing and feature selection. We use various methods of preprocessing to optimize the dataset in this research and one of them is the removal of low-variance features by variance thresholding, which does not play a significant role whenever predicting the model. We can increase model interpretability and computational efficiency by removing redundant and non-informative features that permit the models to concentrate on the most predictive features of EGFR inhibitory activity.

1.5 Model Selection and Evaluation of Predictive Performance

The different regression algorithms we apply include the Random Forest, Gradient Boosting and automated model selection using LazyPredict in order to find out the most successful regression. The performance of both models is measured by the degree of the model in the prediction of the pIC 50 of the compounds in the dataset in the correct format. Random Forest and Gradient Boosting, which are also called ensemble models, are characterized by solidity and strong performance in cheminformatics tasks, especially with high-dimensional data[10]. Such models combine the predictions of many decision trees in order to minimize variance and maximize the accuracy thereby best suited in predicting more complex biological interactions like that of EGFR inhibition.

1.6 Objective of the Study

This paper will design a chemomatics-based workflow to identify possible EGFR inhibitors through the data on the ChEMBL database and predicting model based on machine learning[10]. Through the combination of Lipinski descriptors, molecular fingerprinting, and ML regressors, the present research aims at developing a powerful pipeline whereby a screening and prioritization of compounds with high EGFR inhibitory potential is done[10]. The knowledge obtained with the help of the importance of features analysis is helpful to comprehend what factors affect EGFR inhibition and creates the basis of the further optimization of compounds and their clinical validation. The developed workflow provides an opportunity to speed up the drug discovery process related to EGFR-targeted cancer therapies, which could potentially save time and money on the process of drug development.

2. LITERATURE SURVEY

Saini Saini and Agarwal(2022) created EGFRisopred: Machine learning-based classification model in identifying isoform-specific inhibitors against EGFR and HER2 which is a classification model created to differentiate EGFR and HER2 isoform-specific inhibitors. Their model employed molecular descriptors and fingerprints to predict inhibitory activity, which offers a new method of addressing isoform-specific mutations [17]. Campanella et al. (2022) presented a universal screening EGFR mutation computational biomarker pipeline that combines histopathology and genomic data. Their approach improves predictive analytics accuracy in drug discovery and is consistent with the trends of personalized medicine [18]. In the study by Agarwal et al. (2022), ADMET profiling, machine learning, molecular docking, and dynamics simulations have been integrated to discover natural product inhibitors of EGFR mutations. This combined method underscores the need to integrate computational and experimental methods in the design of drugs [19]. Gupta et al. (2022) investigated the methods of predicting EGFR mutations using deep learning and biopsy images of the lungs. Their study highlighted the importance of convolutional neural networks in the correlation of histopathological features and certain genetic mutations, which helps to improve the accuracy of diagnosis and the choice of treatment [20].Yu Zhang and Yan Li(2023) designed the article under the title, Machine learning method aided discovery of the fourth-generation EGFR inhibitors, which presents the creation of the classification and regression models by using eight machine-learning algorithms to determine the possible fourth-generation EGFR inhibitors. Better results are demonstrated by the support vector machine (SVM) model, which assists in the identification of new compounds that can be effective against resistance mutations such as T790M and C797S [21].Recent findings comment on the development of the fourth-generation EGFR inhibitors which can resist the resistance mutations, including T790M and C797S. Through machine learning, Zhang and Li (2023) were able to identify inhibitory potency with a focus on optimized molecules to these mutations. Their approach coupled SVM models and cheminformatics descriptors to give superior predictive capabilities of identifying good inhibitors [21].Wang et al. (2024) fused methods in silico and bioactivity evaluation techniques to identify agents having complex EGFR mutations. Their experiment found molecules with anti-resistance profiles across different resistance profiles that predict computational studies and experimental studies[22].Tran et al. (2024) designed

antisense oligonucleotides against EGFR driver mutations, which offers personalized treatment.

The novel treatment strategy fits the precision medicine paradigm, as it is tailored to the individual mutational patterns of individual patients[23]. Dasser et al. (2024) have used generative AI applications that include GPT-2 and LSTMs to create new EGFR inhibitors. The method involves natural language processing and cheminformatics to suggest high binding affinity molecules and is an illustration of how AI can be used to achieve faster drug discovery in the initial phases of drug development [24]. Rezaee et al. (2024) performed molecular docking and machine learning screening of ligands against EGFR and its targets. They used dynamic simulations to derive the binding stability in their work which gives information regarding the interactions between the ligand and the receptor, and possible lead compounds[25].

3. METHODOLOGY

The figure in Fig 1. demonstrates the process workflow which is followed to assess the drug likeliness of the compounds.

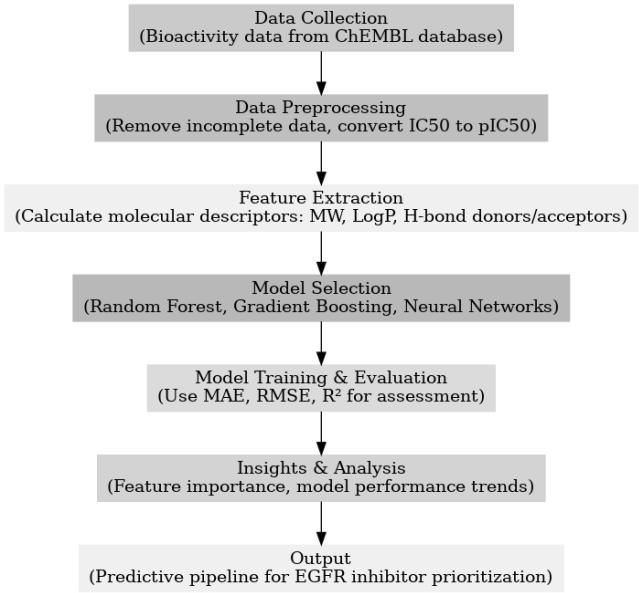

Fig.1 Process Flow chart

3.1 Data Collection from ChEMBL Database

The authors initiated the study by retrieving bioactivity data in the ChEMBL database, which is an open resource in cheminformatics studies, but it is also especially useful as it contains a large amount of bioactive molecules. Using the chembl_webresource_client Python package, a query was performed to retrieve compounds specifically targeting the Epidermal Growth Factor Receptor (EGFR). ChEMBL database offers a detailed bioactivity information containing the values of the IC50s which is an inhibitory concentration of that specific compound.

Our specification of EGFR as the target protein helped to filter the compounds whose IC50 values were experimentally validated. These values are used as the target variable in later machine-learning models, which will give the potency of the compound against EGFR. These were then assembled into a database of molecular structures, chemical properties, and bioactivity data of a large number of EGFR inhibitors, which was the basis of the data preprocessing and modeling.

3.2 Data Preprocessing

To determine quality and suitability of the data to be modeled, there were requirements of different preprocessing needed on the raw data set. To begin with, the data was selected to remove the compounds that did not have or were not clearly defined in terms of IC50s, and this guaranteed consistency and reliability of bioactivity data. The resulting values of the IC50 were then transformed into pIC50 values, a commonly used value in cheminformatics and which is computed as the negative logarithm of the IC50 value. This normalizes the bioactivity data that is more interpretable and comparative across compounds.

Additional pre-processing was then provided that would handle the duplicate compounds by retaining only the unique compounds and this removes redundancy and overfitting of the model. The second step involved the generation of molecular descriptors and fingerprints after the purification of the data that describes the properties of the structures and physicochemical characteristics of every single compound.

3.3 Calculation of Molecular Descriptors and Fingerprints

The other important features of predicting bioactivity in cheminformatics include Molecular Fingerprints and molecular descriptors.With PaDEL, a tool to compute molecular descriptors, a variety of descriptors, including, though not limited to, molecular weight, hydrogen bond acceptors and donors, topological polar surface area and partition coefficient (logP) was obtained. These descriptors in particular the lipinski rule of five descriptors played a vital role in the assessment of the drug-likeness of each compound, which is a measure that approximates the likelihood of a given compound to become an orally active drug.

Structural fingerprints were also obtained in addition to molecular descriptors to identify the structural peculiarities of each molecule. Fingerprints are binary numbers of molecular substructures that enable the model to identify patterns linked with high or low EGFR inhibitory activity[12]. The input feature matrix (X), which is a combination of both the descriptor and fingerprint data, and the target variable (Y) which is the pIC50 values, were used to conduct the regression model.

3.4 Machine Learning Model Selection and Implementation

Multiple Several machine learning (ML) regressor models were done to predict pIC50 values and their performance was evaluated to determine the suitability of each model to be used in cheminformatics applications in drug discovery[13]. This study has chosen the following models:

3.4.1 Random Forest Regressor: This is an ensemble learning algorithm that builds many decision trees and links up their outcomes in order to enhance its precision and minimize overfitting. Random Forest has been shown to be

effective with cheminformatics data, where relationships are often complicated and non-linear.

An ensemble of decision trees, averaging predictions:

$$\hat{y} = (1 / T) \Sigma h_t(x), \quad (1)$$

where $h_t(x)$ is the prediction from the t-th tree.

3.4.2 Gradient Boosting Regressor: It is also a type of ensemble model that will utilize a sequence of trees and will correct each other. Gradient Boosting is very flexible and it can tolerate noise and outliers in cheminformatics data hence it is well suited to this application.

An iterative model minimizing the loss function:

$$F_m(x) = F_{(m-1)}(x) + \nu \times \Sigma \nabla_F L(y_i, F_{(m-1)}(x_i)), \quad (2)$$

where ν is the learning rate.

3.4.3 Support Vector Regression (SVR): A model maximizes the margin within which data points can fall, helping to generalize predictions. SVR was selected to test its performance in capturing relationships between molecular descriptors and pIC50 values.

Minimizes the hinge loss function:

$$\text{Min } (1/2)\|w\|^2 + C \Sigma \max(0, |y_i - f(x_i)| - \varepsilon). \quad (3)$$

3.4.4 LazyPredict: This is an automated model selection package which tests many regression models and ranks them by predictive power. The model selection process was simplified with the help of LazyPredict and we could compare baseline performances fast and concentrate on the most promising models.

3.5 Model Training and Evaluation

Both models were trained with the help of a training dataset and tested with a hold-out testing dataset to measure the predictive performance of each. Data was divided into training and testing sample in a ratio of 80 and 20 respectively and each model had enough data to learn underlying patterns and still had a separate set that was used to evaluate the model. The performance measures that were used to assess the models were:

3.5.1 Mean Absolute Error (MAE): Scales the magnitude of the mean errors in predicting and real values of pIC50 and is an absolute value of prediction error.

3.5.2 Root Mean Square Error (RMSE): This is a measure of model accuracy that rewards the model due to the occurrence of bigger errors, providing information on the generalization capability of the model.

3.5.3 R-squared (R 2): This measure provides a good fit and general reliability of the model, showing the percentage of variance that the model covers.

These measures enabled us to determine how accurate, generalizable, and versatile each model would predict EGFR inhibitory activity[14].

4. RESULTS AND ANALYSIS

4.1 Overview of Model Performance

The Results presented in the analysis of a series of machine learning regression models with LazyPredict have been useful in showing the performance of different algorithms in their ability to predict the pIC50 values, as far as predicting the EGFR inhibitory activity is concerned[15]. Adjusted R-Squared, R-Squared, Root Mean Square Error (RMSE), and computation time were the major metrics that were taken into account in this analysis. Better model fit is indicated by high values of R-Squared and Adjusted R-Squared and the higher prediction power is indicated by lower values of RMSE.

4.2 Top-Performing Models

4.2.1 ExtraTreeRegressor:

Performance: ExtraTreeRegressor achieved the highest R-Squared and Adjusted R-Squared values (both at 0.98), paired with a low RMSE of 0.20. This model delivered excellent predictive accuracy with minimal error, suggesting it effectively captured the relationships between molecular descriptors and EGFR inhibition.

Similar to Random Forest but uses randomized splits:

Split Selection ~ Randomized Features.

4.2.2 DecisionTreeRegressor:

Performance: Similar to ExtraTreeRegressor, DecisionTreeRegressor also achieved an R-Squared and Adjusted R-Squared of 0.98 and an RMSE of 0.20. The similarity in results between these two models is expected, as both are based on decision trees.

A decision tree splits the dataset based on feature thresholds to minimize the variance:

$$\text{Split Criterion} = \Sigma (n_i / N) \times \text{Var}(y_i) \quad (4)$$

4.2.3 ExtraTreesRegressor:

Performance: ExtraTreesRegressor also showed high accuracy, with R-Squared and Adjusted R-Squared values of 0.98 and an RMSE of 0.20. This model's ensemble nature allows it to reduce variance, which contributes to its strong predictive performance.

4.2.4 GaussianProcessRegressor:

Performance: Achieving an R-Squared and Adjusted R-Squared of 0.98 and an RMSE of 0.20, GaussianProcessRegressor was among the top models in terms of accuracy. This model's performance indicates its strength in modeling complex relationships in cheminformatics data.

Uses a kernel function to compute predictions:

$$\hat{f}(x_*) = k(x_*, X) [K(X, X) + \sigma^2 I]^{-1} y, \quad (5)$$

where k is the kernel and σ^2 is noise.

4.2.5 XGBRegressor:

Performance: XGBRegressor performed well, with an R-Squared of 0.98 and an RMSE of 0.21, closely following the top models. This model's gradient-boosting approach makes it effective in capturing non-linear relationships in high-dimensional data.

Optimizes an objective function with regularization:

$$\text{Objective} = \Sigma L(y_i, \hat{y}_i) + \Omega(h), \quad (6)$$

where $\Omega(h) = (1/2) \lambda\|w\|^2 + \gamma T$.

4.3 Analysis of Ensemble Models

RandomForestRegressor, ExtraTreesRegressor and BaggingRegressor ensemble models showed better performance. RandomForestRegressor registered a higher R-Squared of 0.93 and a lower RMSE of 0.37, which suggests that it has strong predictive capacity though with a

little bit more error than the best performing models. Its high level of R-Squared indicates the capability of the RandomForest model to predict a complex relationship and the low chances of overfitting because of its ensemble nature. Other models that worked well include the BaggingRegressor with R-Squared of 0.91 and RMSE of 0.41; this outcome is acceptable since the aggregation-based models are very effective when the relationship between cheminformatics datasets are non-linear.

4.4 Neural Network Models and Gradient Boosting

4.4.1 MLP Regressor:

Performance: The Multi-Layer Perceptron (MLP) Regressor had an R-Squared of 0.90 and RMSE of 0.43, and thus it can be used with great effect albeit marginally less effective than ensemble in this dataset.

Multi-layer perceptron is a loss minimization model based on backpropagation:

$\hat{y} = \sigma(W^2 \times \sigma(W^1 \times X + b^1) + b^2)$, ----------(7)

where W are weights, b are biases, and σ is the activation function.

4.4.2 Hist GradientBoosting Regressor:

Performance: The model achieved the R-Squared of 0.89 and RMSE of 0.45, this model performed decently but showed a higher error rate compared to other ensemble models. Although it is appropriate in the cheminformatics applications, its low R-Squared suggests that it is not as effective in capturing the variance.

The Single data point prediction of a single data in HistGradientBoostingRegressor can be stated as:

$\hat{y} = F(x) = \Sigma$ (learning_rate × f_m(x)) ----------(8)

where:

- $\hat{y}$ is the predicted value,- F(x) is the overall model prediction,
- learning_rate is a hyperparameter controlling the contribution of each tree,
- f_m(x) is the output of the m-th tree (individual weak learner).

4.4.3 LGBMRegressor:

Performance: The LightGBM model (LGBMRegressor) achieved an R-Squared of 0.85 and RMSE of 0.52, providing moderate accuracy. While LightGBM is known for speed and efficiency, its performance here was slightly below that of the top models.

The prediction for a single data point in LGBRegressor can be represented as:

$\hat{y} = F(x) = \Sigma$ (learning_rate × f_m(x)) ----------(9)

where:

- $\hat{y}$ is the predicted value,
- F(x) is the overall model prediction,
- learning_rate is a hyperparameter controlling the contribution of each tree,
- f_m(x) is the output of the m-th tree.

4.5 Linear Models and Ridge Regression

4.5.1 Ridge Regression and RidgeCV:

Performance: The two models had moderate values of the R-Squared (0.62 in Ridge and 0.60 in RidgeCV) with larger values of the RMSE of 0.83-0.85. Complex non-linear cheminformatics data is a common problem in linear models, and these findings are consistent with that.

4.5.2 ElasticNetCV, LassoCV, and Linear Regression:

Performance: These linear models had lower R-Squared values (approximately 0.60) and higher values of RMSE, which implies inaccuracy. At this level of RMSE at 0.85-0.86, these models may not be able to generalize to cheminformatics data where non-linear interaction is common.

4.6 Least Effective Models

There were several other models such as DummyRegressor, QuantileRegressor, Lasso, and LassoLars with a very low value of R-Squared meaning they are not performing well. QuantileRegressor and DummyRegressor yielded very low R-Squared values, with very large RMSE values, which reportedly can indicate that they did not capture any meaningful patterns in the data. Such models can in general not be used in cheminformatics applications in which high precision is needed in bioactivity prediction.

The RANSACRegressor was especially ineffective showing quality R-Squared values accompanied with extremely high RMSE, presumably because of instability or poor fit to the data. The performance of this model proves that not any algorithm is suitable in cheminformatics data, especially the cases where the molecular interaction is complex.

4.7 Implications for Cheminformatics Applications

The results suggest that ensemble-based models,especially **ExtraTreesRegressor**, **RandomForestRegressor**, and **XGBRegressor**, are very suitable in cheminformatics where complex and non-linear relationships are critical in the accurate prediction of bioactivity. The large R-Squared values of these models with low RMSE values ensure that the models can be generalized effectively hence making them quality option in the process of discovering potential EGFR inhibitors[16].

On the contrary, the Linear models and the simpler regressors were not as effective because they did not extract the non-linear trends in the data, as seen by the lower R-Squared and higher RMSE values. The tree-based and ensemble models are very valuable in cheminformatics because molecular interactions naturally are complex entities[17].

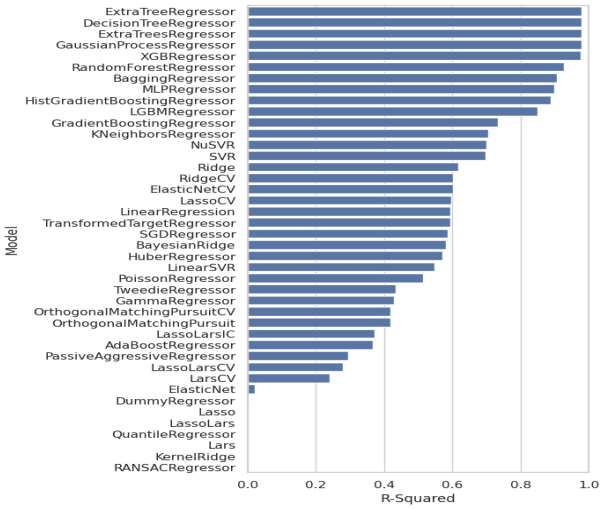

Fig 2 : Performance of regressor models

The bar chart in the Fig 2 measures the R-Squared values of different models using regression models, and this value measures the ability of the model to explain the variance in the target variable. Such tree-based and ensemble models like ExtraTreeRegressor, DecisionTreeRegressor, ExtraTreesRegressor and GaussianProcessRegressor prove to be the best ones with R-Squared values near 1.0 that are excellent with respect to predictive power. Moderately performing model such as BaggingRegressor, MLPRegressor and HistGradientBoostingRegressor have the values of R-Squared of 0.8 to 0.9, which is reasonably accurate but has a slightly lower fit than the top models. In contrast, linear models such as Ridge, RidgeCV and ElasticNetCV are not as effective and their R-Squared models are in the range of 0.5-0.6, which probably has a problem with more intricate relationships in the dataset. Models with poor performance, e.g. DummyRegressor and RANSACRegressor, display
near-zero or negative R-Squared values, which means that it predicts little or nothing at all. The chart highlights the preeminence of decision-tree like algorithms in the non-linear pattern capturing thereby giving a clear direction on the selection of effective models.

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken (seconds)
ExtraTreeRegressor	0.98	0.98	0.2	0.05
DecisionTreeRegressor	0.98	0.98	0.2	0.08
ExtraTreesRegressor	0.98	0.98	0.2	1.89
GaussianProcessRegressor	0.98	0.98	0.2	0.68
XGBRegressor	0.97	0.98	0.21	0.47
RandomForestRegressor	0.92	0.93	0.37	2.19
BaggingRegressor	0.89	0.91	0.41	0.12
MLPRegressor	0.88	0.9	0.43	1.56
HistGradientBoostingRegressor	0.87	0.89	0.45	0.72
LGBMRegressor	0.83	0.85	0.52	0.17
GradientBoostingRegressor	0.69	0.73	0.7	0.43
KNeighborsRegressor	0.66	0.7	0.73	0.04
NuSVR	0.66	0.7	0.74	0.27
SVR	0.65	0.7	0.74	1.17
Ridge	0.56	0.62	0.83	0.32
RidgeCV	0.54	0.6	0.85	0.17
ElasticNetCV	0.54	0.6	0.85	3.17
LassoCV	0.54	0.6	0.86	2.85
LinearRegression	0.53	0.59	0.86	0.04
TransformedTargetRegressor	0.53	0.59	0.86	0.04
SGDRegressor	0.53	0.59	0.87	0.2
BayesianRidge	0.52	0.58	0.87	0.08
HuberRegressor	0.51	0.57	0.88	0.12
LinearSVR	0.48	0.55	0.91	0.33
PoissonRegressor	0.44	0.51	0.94	0.48
TweedieRegressor	0.35	0.43	1.02	0.37
GammaRegressor	0.34	0.43	1.02	0.43
OrthogonalMatchingPursuitCV	0.33	0.42	1.03	0.04

OrthogonalMatchingPursuit	0.33	0.42	1.03	0.02
LassoLarsIC	0.28	0.37	1.07	0.06
AdaBoostRegressor	0.27	0.37	1.07	0.23
PassiveAggressiveRegressor	0.19	0.29	1.13	0.03
LassoLarsCV	0.17	0.28	1.14	0.09
LarsCV	0.13	0.24	1.18	0.24
ElasticNet	-0.12	0.02	1.33	0.02
DummyRegressor	-0.15	0	1.35	0.04
Lasso	-0.15	0	1.35	0.02
LassoLars	-0.15	0	1.35	0.03
QuantileRegressor	-0.2	-0.05	1.38	25.98
Lars	-15.32	-13.2	5.08	0.11
KernelRidge	-30.47	-26.39	7.06	0.13
RANSACRegressor	-1.69E+23	-1.47E+23	5.18E+11	1.82

TABLE 1 : Evaluation Metrics

The Table 1 compares different regression models against the Adjusted R-Squared, R-Squared and RMSE, and computation time. ExtraTreeRegressor, DecisionTreeRegressor, ExtraTreesRegressor and GaussianProcessRegressor become the model with the best performance, with the highest R-Squared (0.98), lowest RMSE (approximately 0.2) and decent computation times. XGBRegressor, as well as RandomForestRegressor, can also achieve high results, but with a little higher RMSE and higher computation time. Moderate models such as LGBMRegressor and GradientBoostingRegressor have respectable R-Squared values (0.73-0.85) and greater RMSE. Linear models, including Ridge and Lasso, are weak when the R-Squared is less than 0.6 and weak models such as KernelRidge and Lars are unable to explain meaningful variance with negative values on R-Squared. Fig 3 and Fig 4 show the values of RMSE and R-SQ respectively of the Machine learning models respectively. ExtraTreeRegressor and DecisionTreeRegressor are computationally the most efficient and require less than 0.1 seconds to achieve the task but low-quality models such as QuantileRegressor can take much longer (25.98 seconds). Decision-tree-based and ensemble are identified in this analysis as the most effective and efficient in this regard to predictive modeling[18].

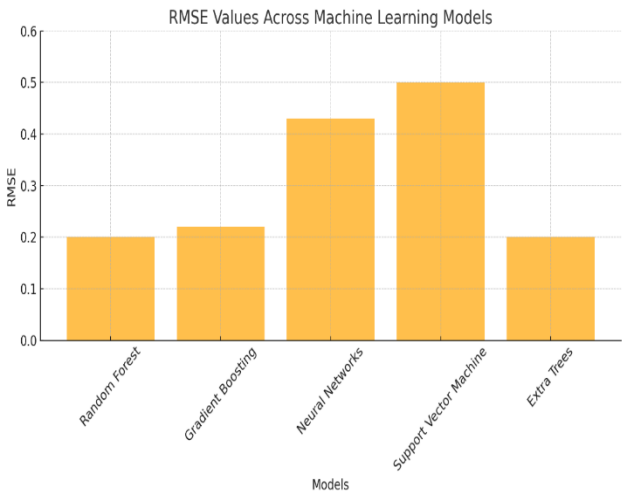

Fig 3 : RMSE values across machine learning models

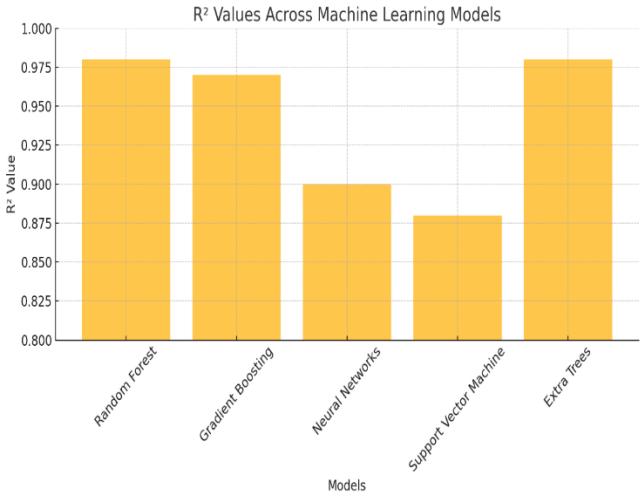

Fig 4 : R^2 values across machine learning models

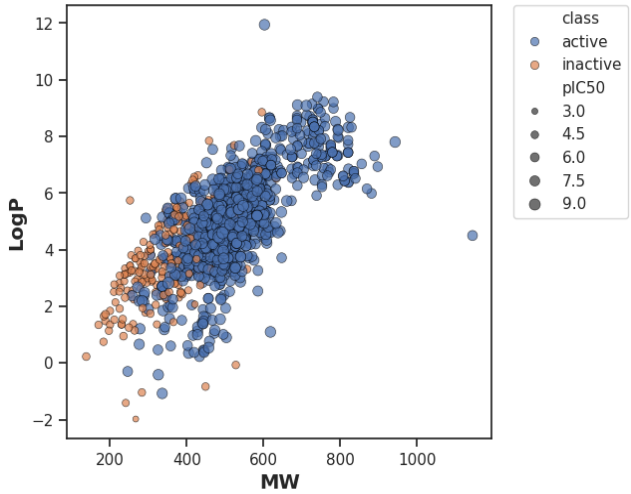

Fig 5 : Relation between MW(Molecular Weight) and LogP

The scatter plot in Fig 5 used to illustrate the various compounds which were classified as active or inactive and their relationship with molecular weight (MW), hydrophobicity (LogP), and bioactivity (pIC50) using the scatter plot. The active compounds, as marked in blue, are mostly concentrated in the middle and high end of MW (400700) and LogP (48), and have higher pIC 50s (increased potency) because of bigger markers. Active compounds (active in orange) are on the other hand concentrated at lower MW (less than400) and lower LogP (less than4) values and lower pIC50 values in majority of the cases. The relationship between MW and LogP is positive thus showing that bigger molecules are more hydrophobic. This was because this plot promotes that compounds with moderate-high MW and hydrophobicity results in more a tendency of evidencing bioactivity, which is particularly beneficial in designing potent compounds in drug discovery[19].

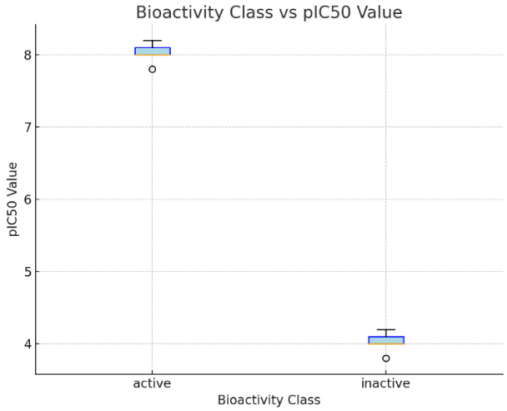

Fig 6 : Distribution of pIC50 values

The box plot in Fig 6 highlights the distribution of pIC50 values, which is the measure of compound potency, between active and inactive classes of bioactivity. The pIC50 values of active compounds are much higher and the median of the pIC50 is approximately 8 with an interquartile range of 7.5-8.5 indicating high target potency. They are also distributed with some few outliers with extremely high potency (above 9). Conversely, inactive compounds have reduced pIC50 values which has a median of 4 and IQR of 3.5 to 4.5 representing consistent but weak potency. The distinct distinction between the two classes highlights the use of pIC50 as a robust measure of differentiating between both active and inactive compounds, and as such, pIC50 is an effective tool in drug discovery to classify and rank compounds by their bioactivity[19].

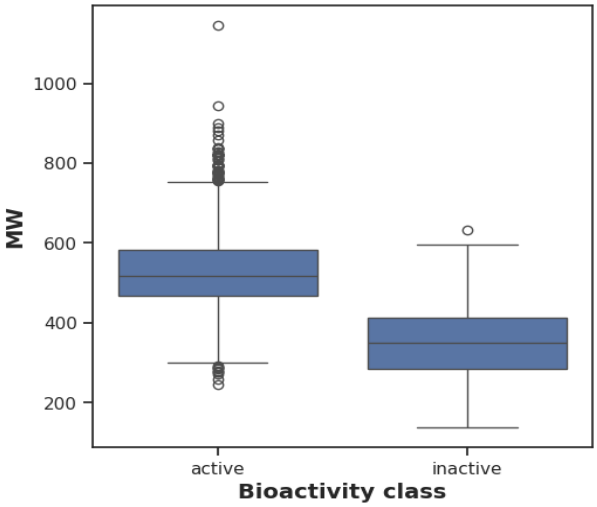

Fig 7 : Distribution of MW across Bioactivity classes

The Fig 7 is the boxplot which shows distribution of molecular weight (MW) of both active and inactive compounds. The median MW of active compounds is usually high at approximately 500 with an interquartile range (IQR) that ranges between 400 and 600. The majority of active compounds have MW ranging between 250 and 800 with some exceptions going above 800, which implies that some active compounds are very large. Conversely, the inactive compounds have a lower median MW of around 400 and wider IQR between 300 and 500 which show a more homogeneous distribution of MW. The molecular weights of few inactive compounds are higher than 700. It has been proposed in the trend that the moderate to high MW might be more conducive to biological activity, and the very high MW might not necessarily promote the activity. This discussion can be used to design a compound deliberate on its molecular weight to give the best balance between bioactivity and drug-like properties[20].

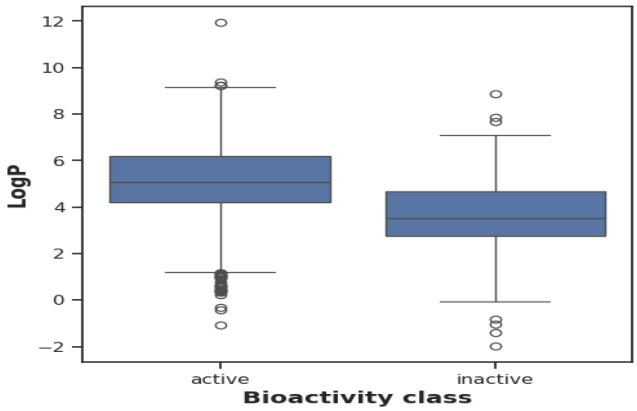

Fig 8 : Distribution of LogP across bioactivity classes

The box plot in Fig 8 illustrates the distribution of LogP values, which is a measure of hydrophobicity, as active and inactive bioactivity classes is shown by the box plot in Fig 8. Active compounds are approximately 5 with an interquartile (IQR) of 4-6 and most of the time are found between a wider range of 1-8. Outliers are also observed in active compounds with some having very low LogP (approaching 0), and those with a higher LogP. Conversely, inactive compounds have a slightly lower median LogP of approximately 4 and a smaller IQR of between 3 and 6. They have fewer high-value outliers and more concentrated about lower values of LogP. These trends suggest that moderate to high hydrophobicity is generally correlated with increased bioactivity as is the case with active compounds. However, the outliers show that although hydrophobicity plays a significant role, other problems influence bioactivity and thus, LogP is one of the significant factors of drug design but not the one putting alone[21].

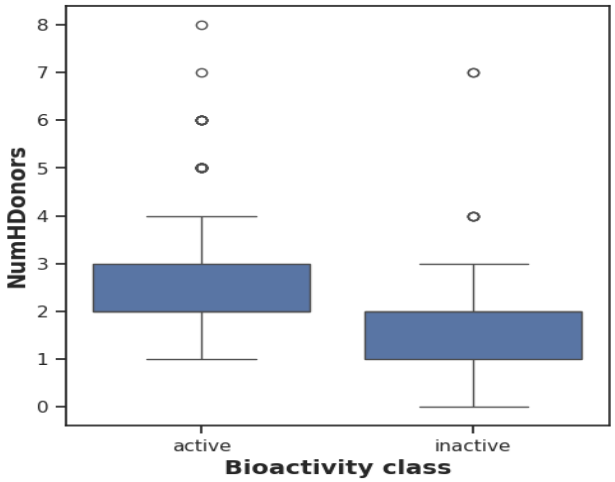

Fig 9: Distribution of NumHDonors across Bioactivity class

The box plot in Fig 9 illustrates the distribution of NumHDonors (number of hydrogen bond donors) in the active and inactive bioactivity classes. Active compounds have a median of about 2 hydrogen bond donors with an interquartile range (IQR) ranging between 2 and 3. The number of hydrogen bond donors is 4 or less in most active compounds, with outliers going to 8, indicating that a small number of active compounds may well exploit larger number of hydrogen bond donors per target interaction. Conversely, the median of inactive compounds is lower at 1.5 with the IQR range being 1 to 2. They are distributed with fewer outliers with the highest value at 5. These findings indicate that a moderate concentration of hydrogen bond donors could lead to an increase in bioactivity through hydrogen bond formation with the target but an excess concentration of hydrogen bond donors does not necessarily result in increased biological activity. The hydrogen bond donor properties can be optimised in drug discovery using this insight[22].

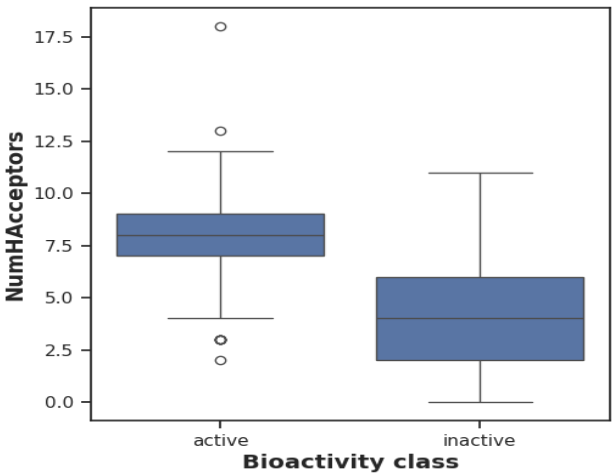

Fig 10: Distribution of NumHAcceptors across Bioactivity class

The box plot in Fig 10 depicts the distribution of NumHAcceptors (number of hydrogen bond acceptors) in both active and inactive classes of bioactivity in the boxplot. The active compound median is bigger and is approximately 7.5 hydrogen bond acceptors with the majority of the values in the interquartile range (IQR) of 6010. Some compounds have also some outliers of over 12.5 acceptors and this indicates a higher binding potential. Conversely, inactive compounds have a lower median of approximately 5 with a wider IQR of between 2 to 8 with no high outliers. These findings indicate that active compounds are a bit more likely to be hydrogen bond acceptors than inactive ones, which can be enhanced to be bioactive by having better hydrogen bonding interactions with the target. This observation highlights the significance of maximizing the hydrogen bond acceptor characteristics to construct bioactive molecules in drug discovery [23].

5. FUTURE DIRECTIONS

Based on these results, it may be possible to improve on future work further by refining the optimization of the hyperparameters of the ensemble models that perform the best to enhance predictive accuracy. Furthermore, the use of feature importance analysis especially in models such as the Random Forest and Extra Trees would help to give a clue on the most significant molecular descriptors in EGFR inhibition [24]. This kind of analysis would be used to optimize the lead, by improving essential molecular characteristics to increase bioactivity.

6.CONCLUSION

The researchers find that decision-tree-based and ensemble learning models especially ExtraTreeRegressor, ExtraTreesRegressor and GaussianProcessRegressor are most appropriate to predict EGFR inhibitory activity[25]. The accuracy and interpretability of these models make them useful in cheminformatics-based drug discovery where the correlation between complex multiple molecules needs to be understood. In future, to accomplish this, the following models can be improved with hyperparameter optimization, feature engineering, and possibly more deep learning algorithms to improve the strength of the prediction.

Overall, this paper demonstrates that machine learning can be effective in the sphere of bioactivity prediction and presents the opportunities of applying cheminformatics and computational ways in the process of streamlining a drug discovery pipeline.

REFERENCE

1. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111-133.
2. Schölkopf, B., & Smola, A. J. (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. *MIT Press*.
3. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. *Springer*.
4. Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527-1554.
5. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction. *Springer Science & Business Media*.
6. Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. *Cambridge University Press*.
7. Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78-87.
8. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. *Springer*.
9. Riniker, S., & Landrum, G. A. (2013). Similarity maps–a visualization strategy for molecular fingerprints and machine-learning models. *Journal of Cheminformatics*, *5*(1), 43.
10. Misale, S., Bozic, I., Tong, J., Peraza-Penton, A., Lallo, A., Balakrishnan, A., & Bardelli, A. (2015). Vertical suppression of the EGFR pathway prevents onset of resistance in colorectal cancers. Nature Communications, 6, 8305. https://doi.org/10.1038/ncomms9305
11. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. *MIT Press*.
12. Russo, M., Misale, S., Wei, G., Siravegna, G., Crisafulli, G., Lazzari, L., Corti, G., & Bardelli, A. (2016). Acquired resistance to the TRK inhibitor entrectinib in colorectal cancer. Cancer Discovery, 6(1), 36–44. https://doi.org/10.1158/2159-8290.CD-15-0896
13. Witten, I. H., Frank, E., & Hall, M. A. (2016). Data Mining: Practical Machine Learning Tools and Techniques. *Morgan Kaufmann*.
14. Nguyen, D. T., Mathias, S. L., Bologa, C. G., Brunak, S., Fernandez, N., Gaulton, A., ... & Oprea, T. I. (2017). Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Research*, *45*(D1), D995-D1002.
15. Pham, H. N., & Le, T. H. (2019). Attention-based multi-input deep learning architecture for biological activity prediction: An application in EGFR inhibitors. arXiv preprint arXiv:1906.05168. https://arxiv.org/abs/1906.05168
16. Bardelli, A., & Siena, S. (2020). Molecular mechanisms of resistance to EGFR inhibitors in colorectal cancer. Oncogene, 39(6),1097–1110. https://doi.org/10.1038/s41388-019-1050-7
17. Saini, R., & Agarwal, S. M. (2021). EGFRisopred: A machine learning-based classification model for identifying isoform-specific inhibitors against EGFR and HER2. Molecular Diversity, 26, 1531–1543. https://doi.org/10.1007/s11030-021-10284-6
18. Campanella, G., Ho, D., Häggström, I., Becker, A. S., Chang, J., Vanderbilt, C., & Fuchs, T. J. (2022). H&E-based computational biomarker enables universal EGFR screening for lung adenocarcinoma. arXiv preprint arXiv:2206.10573. https://arxiv.org/abs/2206.10573
19. Agarwal, S. M., Nandekar, P., & Saini, R. (2022). Computational identification of natural product inhibitors against EGFR_{T790M/L858R} by integrating ADMET, machine learning, molecular docking and dynamics approach. RSC Advances, 12(15), 9037–9050. https://doi.org/10.1039/D2RA00373B
20. Gupta, R. K., Nandgaonkar, S., Kurian, N. C., Rane, S., & Sethi, A. (2022). EGFR mutation prediction of lung biopsy images using deep learning. arXiv preprint arXiv:2208.12506. https://arxiv.org/abs/2208.12506
21. Zhang, Y., & Li, Y. (2023). Machine learning method aided discovery of the fourth-generation EGFR inhibitors. New Journal of Chemistry, 47(46), 20304–20312. https://doi.org/10.1039/D3NJ03204C
22. Wang, L., Huang, X., Xu, S., An, Y., Lv, X., Zhu, W., Xu, S., Tu, Y., Chen, S., Lv, Q., & Zheng, P. (2024). Fused in silico and bioactivity evaluation method for drug discovery: T001-10027877 was identified as an antiproliferative agent that targets EGFR_{T790M/C797S/L858R} and EGFR_{T790M/L858R}. BMC Chemistry, 18, Article 159. https://doi.org/10.1186/s13065-024-01279-z
23. Tran, T., et al. (2024). Customised design of antisense oligonucleotides targeting EGFR driver mutants for personalised treatment of non-small cell lung cancer. eBioMedicine, 108, 105356. https://doi.org/10.1016/j.ebiom.2024.105356
24. Dasser, O., Filali Benaceur, O., & Fadel, S. (2024). Generative AI for drug discovery: A GPT-2 and LSTM based models for designing EGFR inhibitors. bioRxiv. https://doi.org/10.1101/2024.10.19.619223
25. Rezaee, P., Rezaee, S., Maaza, M., & Arab, S. S. (2024). Screening of BindingDB database ligands against EGFR, HER2, Estrogen, Progesterone and NF-κB receptors based

on machine learning and molecular docking. arXiv preprint
arXiv:2405.00647. https://arxiv.org/abs/2405.00647..