Machine Learning-Based Prediction of EGFR Bioactivity Using Molecular Fingerprints

Kandula Siri Chandana[1], Vanitha Kakollu[2]

[1]PG Student, Department of Computer Science, GSS, GITAM Deemed to be University

[2]Assistant Professor, Department of Computer Science, GSS, GITAM Deemed to be University

DOI: https://doi.org/10.51244/IJRSI.2026.1304000079

Received: 04 April 2026; Accepted: 10 April 2026; Published: 01 May 2026

ABSTRACT

The process of drug discovery involves a number of factors and can be described as complicated, time-taking and costly. EGFR has become one of the main targets for further investigation in oncological diseases research. To discover new medicines, it is necessary to discover active chemicals against EGFR. This work proposes the use of machine learning to predict bioactivity based on molecular fingerprints extracted from the SMILES string of a compound. The used dataset contains data from the ChEMBL database. The dataset was preprocessed into binary classes of bioactive molecules. We implemented a variety of machine learning models such as Random Forest, Support Vector Machine, Logistic Regression, Gradient Boosting, and XG Boost. The best performance among all tested models was provided by Random Forest. The obtained accuracy was 87%. The implementation of the model was done using Streamlit web framework.

Keywords: EGFR Bioactivity, Machine Learning, Molecular Fingerprints, Drug Discovery, Random Forest, ChEMBL, Bioactivity Prediction.

INTRODUCTION

The process of drug discovery is complicated and time-consuming, requiring rigorous testing and substantial financial investment. Conventional methods involve compound screening through experimental means; however, such approaches become inefficient when applied to large datasets of chemicals. Epidermal Growth Factor Receptor (EGFR) plays a vital role in oncology as its dysfunctional state is linked to tumor development. Thanks to recent advances in computing, machine learning algorithms prove to be powerful tools that can facilitate rapid discovery of drugs during early stages. Chemical compounds can be encoded using SMILES notation and then converted into numerical descriptors like molecular fingerprints. Machine learning algorithms use such descriptions for predicting biological activity. This paper aims at designing an algorithm capable of making predictions regarding EGFR bioactivity based on molecular fingerprints.

Related Work

There have been various studies on the implementation of machine learning in predicting drugs and bioactivity. There has been a demonstration of very good performance from algorithms such as Random Forest and Support Vector Machine in dealing with chemical data with a very high dimensionality. There is a heavy use of molecular fingerprinting including Morgan fingerprints in describing chemical structures due to their relevance.

Gradient Boosting has been another technique that has been used with great success in improving the accuracy of predictions through the combination of several models. Bioactivity data in public databases such as ChEMBL can be useful in developing models for prediction. Nevertheless, there has been little consideration for implementing such models into convenient software applications for making predictions.

Problem Statement

The drug discovery process is very costly and lengthy, with extensive experimental verification. The traditional

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

approach to screening is not effective in handling large sets of chemical data. Prediction of EGFR bioactivity is difficult due to complex structures. Limitations of accuracy in existing computational methods. Complex process of converting chemical data to meaningful features. Lack of predictive systems for easy predictions. Need for an effective machine learning prediction system.

Dataset Description

The utilized dataset comes from the ChEMBL database that stores bioactivities of chemical compounds that act on the EGFR. The dataset consists of SMILES and IC50 values.For data preprocessing, the process starts with the removal of all missing values. Then all the IC50 values in nanomoles are filtered out and duplicates removed. Afterwards, IC50 values are converted to pIC50 values. Compounds in the dataset are classified into active or inactive based on their pIC50 values.

Sample Data:

	A	B	C
1	canonical_smiles	pIC50	activity
2	C=CC(=O)Nc1cc(Nc2ncc(C)c(-c3cc(F)c4nc(C)n(C(C)C)c4c3)n2)c	7.60206	1
3	C=CC(=O)Nc1cc(Nc2nccc(-c3cc(F)c4nc(C)n(C(C)C)c4c3)n2)c(O(	9	1
4	C=CC(=O)Nc1cc(Nc2nccc(-c3cc(F)c4nc(C)n(C(C)C)c4c3)n2)c(O(	7.60206	1
5	CN(c1cccc(Br)c1)c1nc(N)nc2[nH]c(Cc3ccccc3)cc12	4.27491	0
6	CN(c1cccc(Br)c1)c1nc(N)nc2c1cc(Cc1ccccc1)n2C	3.59585	0
7	C=CC(=O)Nc1cccc(Nc2nc(Nc3ccc(N4CC5CCC(C4)N5C(C)=O)cc3	5.20761	0
8	COc1cc(-c2nn(C(C)C)c3ncnc(N)c23)ccc1F	5.41341	0
9	CC(C)n1nc(-c2ccc(F)c(O)c2)c2c(N)ncnc21	5.85387	0
10	C[C@@H](CN(C)C(=O)CO)Oc1cccc2ncnc(Nc3ccc(OCc4ccccn4)	6.71897	1

PROPOSED METHODOLOGY

The methodology proposed here involves the application of structured machine learning steps to achieve the project aims. First, the data preprocessing step will be conducted to clean the dataset. Then, the chemical compounds in the SMILES format are transformed into molecular fingerprints using RDKit.

The fingerprints obtained above will be used as the input feature in machine learning models. The stratified sampling technique will be employed to split the dataset into the training and test sets. Then models will be trained and evaluated, and the best-performing one will be selected. At last, the model will be put into action using the web interface.

Algorithms Used

The following machine learning algorithms are used in this study:

- Random Forest
- Logistic Regression
- Support Vector Machine (SVM)
- Gradient Boosting Machine (GBM)
- XGBoost

These algorithms are selected to compare different learning approaches and identify the most effective model for bioactivity prediction.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Model Selection

Model selection is done based on evaluation metrics like accuracy, F1 score, and ROC AUC. In all the models, the Random Forest Classifier gave the best results since it could deal with high dimensional data and avoid overfitting by leveraging ensemble learning. Random Forest is appropriate for this project because it can detect complicated relationships between molecules and give consistent outputs. Thus, it is adopted as the final model to use.

Implementation

Data Loading:

```
# Load data

X = np.load("X_fingerprints.npy")

y = np.load("y_labels.npy")
```

Preprocessing:

```
# Train-test split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
```

Model Training:

```
# Train Random Forest

rf = RandomForestClassifier(n_estimators=300,   random_state=42,    n_jobs=-1)

rf.fit(X_train, y_train)
```

Prediction:

```
y_pred = rf.predict(X_test)

y_prob = rf.predict_proba(X_test)[:, 1]
```

Sample code:

```
import numpy as np

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report, roc_auc_score, roc_curve

import joblib

import matplotlib.pyplot as plt

# Load data

X = np.load("X_fingerprints.npy")

y = np.load("y_labels.npy")
```

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

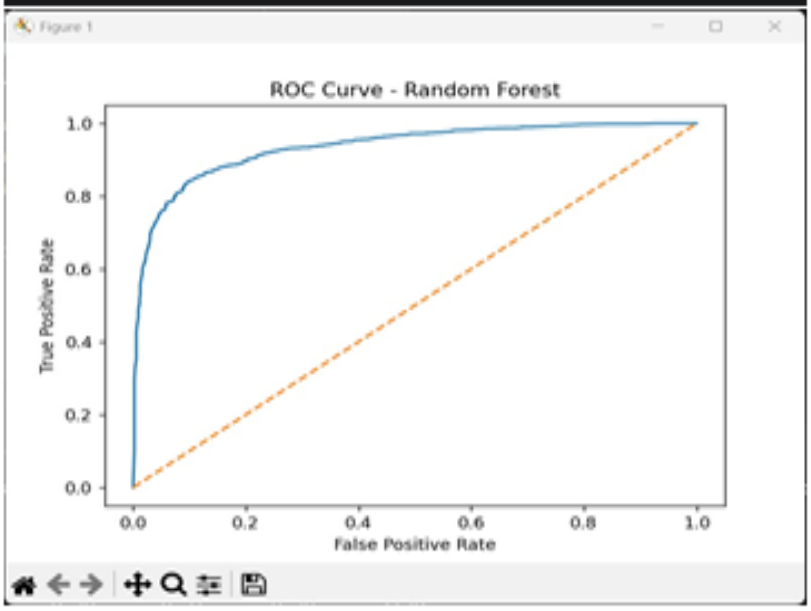
```
print("X shape:", X.shape)
print("y shape:", y.shape)
# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
# Train Random Forest
rf = RandomForestClassifier(n_estimators=300,  random_state=42,  n_jobs=-1)
rf.fit(X_train, y_train)
# Predictions
y_pred = rf.predict(X_test)
y_prob = rf.predict_proba(X_test)[:, 1]
# Evaluation
acc = accuracy_score(y_test, y_pred)
roc = roc_auc_score(y_test, y_prob)
cm = confusion_matrix(y_test, y_pred)
print("\n✅ Random Forest Results")
print("Accuracy:", acc)
print("ROC-AUC:", roc)
print("\nConfusion Matrix:\n", cm)
print("\nClassification Report:\n", classification_report(y_test, y_pred))
# ROC Curve
fpr, tpr, _ = roc_curve(y_test, y_prob)
plt.figure()
plt.plot(fpr, tpr)
plt.plot([0, 1], [0, 1], linestyle="--")
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC Curve - Random Forest")
plt.show()
# Save model
joblib.dump(rf, "random_forest_egfr.pkl")
print(" Model saved as random_forest_egfr.pkl")
```

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue IV April 2026

Results and Output

Random Forest model yielded an accuracy of about 87% when tested on the test set. Random Forest model showed good results when classifying compounds as being either active or inactive, evidenced by the high value of the F1-score. The confusion matrix also suggests that there was an equal number of correctly classified examples.Ensemble approaches, such as Random Forest and Gradient Boosting models, have proven to be superior compared to other methods used. These results support the validity of the methodology and highlight its usefulness as an auxiliary tool in drug discovery.

```
X shape: (13286, 2048)
y shape: (13286,)

✅ Random Forest Results
Accuracy: 0.866064710308502 6
ROC-AUC: 0.935201257210854 6

Confusion Matrix:
 [[ 672  188]
 [ 168 1630]]

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.78      0.79       860
           1       0.90      0.91      0.90      1798

    accuracy                           0.87      2658
   macro avg       0.85      0.84      0.85      2658
weighted avg       0.87      0.87      0.87      2658
```



CONCLUSION

In this paper, we present a method of predicting EGFR activity based on machine learning and molecular fingerprints. The use of the proposed machine learning algorithm allows one to achieve good accuracy rates.

The implementation of the machine learning models into a Streamlit application makes it possible to get predictions instantly. This study shows how machine learning techniques can help optimize the process of discovering new drugs.

REFERENCES

1. D. Mendez et al., “ChEMBL: Towards direct deposition of bioassay data,” Nucleic Acids Research, vol. 47, no. D1, pp. D930–D940, 2019.
2. G. Landrum, “RDKit: Open-source cheminformatics software,” 2023. [Online]. Available: https://www.rdkit.org
3. F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
4. L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
5. J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.
6. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in Proc. ACM SIGKDD, 2016, pp. 785–794.
7. C. Cortes and V. Vapnik, “Support-vector networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
8. D. Rogers and M. Hahn, “Extended-connectivity fingerprints,” Journal of Chemical Information and Modeling, vol. 50, no. 5, pp. 742–754, 2010.
9. A. Lavecchia, “Machine-learning approaches in drug discovery: Methods and applications,” Drug Discovery Today, vol. 20, no. 3, pp. 318–331, 2015.

ABOUT AUTHOR

	Kandula Siri Chandana, pursuing Master of Data Science, Department of Computer Science, GSS, GITAM (Deemed to be University), Visakhapatnam. Her area of interest in Machine Learning.
	Dr Vanitha Kakollu is currently working as Assistant Professor in the Department of Computer Science, GIS, GITAM (Deemed to be University). Her main areas of research include Image Processing, Data Mining and Machine Learning.