Research Article
Journal of Intelligent Systems: Theory and Applications 9 (2026) 1753085
DOI: 10.38016/jista.1753085

Statistical Learning-Based Prediction of Estrogen Receptor Alpha (ERα) Inhibitor Activities

Fatma Karateke[1*], Bilge Özlüer Başer[2], Ayça Çakmak Pehlivanlı[3]

[1,2,3] Department of Statistics, Faculty of Science and Letters, Mimar Sinan Fine Arts University, İstanbul, Türkiye

20232107002@ogr.msgsu.edu.tr, bilge.baser@msgsu.edu.tr, ayca.pehlivanli@msgsu.edu.tr

Abstract
Estrogen receptor alpha (ERα) is a protein that plays a role in processes such as cell growth and proliferation; however, it has become an important research topic due to its overexpression in 70% of breast cancers. ERα inhibitors stop the growth of cancer cells by blocking the activity of this protein. Traditional drug discovery methods are disadvantageous in terms of time and cost. Various approaches exist in the literature for the discovery of ERα inhibitors. Therefore, a machine learning-based Quantitative Structure-Activity Relationship (QSAR) approach was preferred in this study for the discovery of ERα inhibitors. In this study, a machine learning-based QSAR approach was preferred due to its capacity to extract structure-activity relationships from large chemical datasets and its ability to provide high-throughput screening opportunities. This method enables the prediction of biological activities by expressing the chemical structural properties of molecules through numerical descriptors. The ChEMBL206 target identifier was selected as the data source due to its widespread use, high data quality, and the opportunity for comparison with previous studies. The obtained molecules were classified according to their IC50 values, and their chemical space distributions were analyzed using Lipinski rules. Subsequently, 3153 molecular descriptors were calculated for 3053 molecules using the PADEL program. Feature importance analysis revealed that fingerprints such as PubchemFP667 and PubchemFP527, as well as APC2D atom pair descriptors that stood out in the LightGBM model, played critical roles in ERα inhibition. The developed models demonstrated superior performance with accuracy above 94%, sensitivity around 90%, specificity above 95%, and AUC values above 0.97. This study contributes to the efficiency of the drug discovery process by demonstrating that the activity of ERα inhibitors can be predicted with high accuracy rates.

Keywords: Estrogen receptor alpha (ERα), Classification-based-QSAR, Machine learning, Feature importance, Inhibitor activity prediction

Östrojen Reseptör Alfa (ERα) İnhibitörlerinin Aktivitelerinin İstatistiksel Öğrenme ile Tahmini

Öz
Östrojen reseptör alfa (ERα), hücre büyümesi ve çoğalması gibi süreçlerde görev alan bir proteindir; ancak meme kanserlerinin %70'inde aşırı bulunması nedeniyle önemli bir araştırma konusudur. ERα inhibitörleri bu proteinin aktivitesini engelleyerek kanser hücrelerinin büyümesini durdurur. Geleneksel ilaç keşif yöntemleri zaman ve maliyet açısından dezavantajlıdır. ERα inhibitörlerinin keşfi için literatürde çeşitli yaklaşımlar mevcuttur. Bu nedenle, çalışmada ERα inhibitörlerinin keşfi için makine öğrenmesi tabanlı Kantitatif Yapı-Aktivite İlişkisi (Quantitative Structure-Activity Relationship - QSAR) yaklaşımı tercih edilmiştir. Bu çalışmada, geniş kimyasal veri setlerinden yapı-aktivite ilişkilerini çıkarma kapasitesi ve yüksek verimli tarama imkanı sunması nedeniyle makine öğrenmesi tabanlı QSAR yaklaşımı tercih edilmiştir. Bu yöntem, moleküllerin kimyasal yapısal özelliklerini sayısal tanımlayıcılarla ifade ederek biyolojik aktivitelerini tahmin etmeye olanak sağlar. Veri kaynağı olarak, yaygın kullanımı, yüksek veri kalitesi ve önceki çalışmalarla karşılaştırma imkanı sunması nedeniyle ChEMBL206 hedef tanımlayıcısı seçilmiştir. Elde edilen moleküller IC50 değerlerine göre sınıflandırılmış, Lipinski kurallarıyla kimyasal uzay dağılımları analiz edilmiştir. Ardından PADEL programı ile 3053 molekül için 3153 moleküler tanımlayıcı hesaplanmıştır. Özellik önemi analizinde, PubchemFP667 ve PubchemFP527 gibi parmak izlerinin ve LightGBM modelinde öne çıkan APC2D atom çifti tanımlayıcılarının ERα engellemesinde kritik rol oynadığı görülmüştür. Geliştirilen modeller %94'ün üzerinde doğruluk, %90 civarında duyarlılık, %95'in üzerinde özgüllük ve 0.97'nin üzerinde AUC değeriyle üstün performans göstermiştir. Bu çalışma, yüksek doğruluk oranıyla ERα inhibitörlerinin aktivitesinin tahmin edilebileceğini göstererek ilaç keşif sürecinin verimliliğine katkı sağlamaktadır.

Anahtar Kelimeler: Östrojen reseptör alfa (ERα), Sınıflandırma tabanlı QSAR, Makine öğrenmesi, Özellik önemi, İnhibitör aktivitesi tahmini

* Corresponding Author.
E-mail: 20232107002@ogr.msgsu.edu.tr

Recieved : 29 Jul 2025
Revision : 3 Oct 2025
Accepted : 29 Jan 2026

1. Introduction

Modern drug development processes are undergoing a significant transformation with the advancement of technology. The drug discovery process consists of complex stages that require high costs and long time periods (Arciniegas et al., 2000). Laboratory experiments used in traditional approaches are fundamentally examined in two categories: in vitro experiments conducted outside living organisms and in vivo tests performed directly on living organisms (Gümüştaş & Çakmak Pehlivanlı, 2021).

In recent years, with the developments in the pharmaceutical industry, traditional trial-and-error methods that take an average of 12 years and reach a cost of 1.8 billion USD have begun to be replaced by faster and more economical intelligent drug development methods that are based on statistical thinking and supported by mathematics and computer sciences (Shaker et al., 2021). Particularly, statistics-based machine learning algorithms and artificial intelligence methods are increasingly being used in solving complex problems that are difficult to solve with traditional methods in many important fields such as computer science, medicine, finance, and engineering, thanks to the analysis and interpretation of large datasets (LeCun et al., 2015).

These methods, called 'in-silico' in the scientific world, have the ability to predict the potential effects of candidate drug molecules and provide valuable preliminary information before proceeding to laboratory tests. The advantages provided by a correct in-silico approach are multifaceted: Firstly, it guides researchers on which molecules should be subjected to laboratory tests. Additionally, it enables more effective design of experiments to be conducted, thereby reducing the use of experimental animals, allowing the optimal concentrations of molecules to be tested to be determined in advance, and providing significant savings in both time and cost (Gümüştaş & Çakmak Pehlivanlı, 2021).

The importance of these technological developments and in-silico approaches becomes distinctly evident, particularly in the development process of estrogen receptor alpha (ERα) inhibitors used in breast cancer treatment. The development of estrogen receptor alpha (ERα) inhibitors constitutes a critical example demonstrating the importance of modern drug technologies and in-silico approaches. ERα is an important member of the nuclear receptor family in the human body and plays a central role in the regulation of fundamental cellular processes such as cell growth, differentiation, and proliferation (Jensen & Jordan, 2003; Deroo & Korach, 2006).

In approximately 70% of breast cancer cases, excessive production of ERα protein is observed, and this situation causes uncontrolled proliferation of cancer cells (Ali & Coombes, 2002; Musgrove & Sutherland, 2009). This finding has made ERα one of the priority molecular targets in breast cancer treatment (Jordan, 2003; Anderson, 2002). Modern drug development technologies and in-silico approaches play an important role in the design and optimization of ERα inhibitors, which enables the development of more effective and selective treatment options (Huang et al., 2010; Sliwoski et al., 2013).

In the development process of ERα inhibitors, modern computer technologies come into play to accelerate drug discovery and development studies. This process begins primarily with the collection of molecular structures and their activity values from reliable data sources. ChEMBL is an open-access database developed and managed by the European Bioinformatics Institute, which provides researchers with experimental data on more than 2 million compounds and their biological activities (Gaulton et al., 2017). On the molecules obtained from this database, drug-likeness parameters known as Lipinski's 'Rule of Five' are first examined. These rules (molecular weight, LogP, hydrogen bond donor and acceptor numbers) are fundamental criteria used in evaluating the potential of a molecule to act as a drug when taken orally (Lipinski, 2004). Analysis of Lipinski rules reveals the distributions of active and inactive molecules in chemical space, enabling the early-stage identification of potential candidates with drug-like properties. These rules are widely used, particularly in predicting the absorption and permeability properties of orally administered drugs, helping to filter molecules in the early stages of the drug development process (Lipinski et al., 2001). Following Lipinski's analysis, the PADEL (Pharmaceutical Data Exploration Laboratory) program is used for more comprehensive structural characterization of molecules. PADEL is an open-source software that makes important contributions to the drug design process by calculating molecular descriptors and fingerprints (Yap, 2011). Thousands of different molecular descriptors can be calculated with this program, which represents the structural details of molecules more deeply.

Following this comprehensive molecular data analysis, statistical learning methods come into play to accelerate drug discovery and development studies in the development process of ERα inhibitors. Prediction models developed on PADEL descriptors perform active-inactive classification to determine which structural features have the strongest relationship with activity. In recent years, ensemble learning algorithms (such as Random Forest, XGBoost, LightGBM) have shown significant success in these analyses (Al-Thanoon et al., 2019; Qasim & Algamal, 2018). These algorithms provide effective results in determining important structural features by evaluating thousands of properties of molecules simultaneously Through feature importance analysis, valuable information is obtained about the structural features to be targeted in the optimization of ERα inhibitors and the design of new potential inhibitors. This approach offers a faster and more economical evaluation opportunity compared to traditional experimental methods for complex biological targets like ERα (Carracedo-Reboredo et al., 2021).

In this study, a comprehensive approach has been adopted in predicting the activity of ERα inhibitors. Molecules obtained from the ChEMBL206 database were first classified as active and inactive according to their IC50 (half-maximal inhibitory concentration) values.

IC50 is an important pharmacological parameter that expresses the concentration required for an inhibitor to reduce the activity of its target protein by 50% (Sebaugh, 2011). The lower this value, the higher the inhibitor activity of the relevant molecule (Zhang et al., 2000).

To evaluate the drug likeness of the classified molecules, their physicochemical properties were analyzed using Lipinski's Rule of Five, and their distribution within the chemical space was examined. For detailed characterization of the structural properties of molecules, the PADEL program was used, and a total of different molecular descriptors were calculated for all molecules in the dataset.

2. Materials and Methods

In this study, a systematic and comprehensive approach was adopted for predicting the activity of ERα inhibitors. The research methodology includes data characterization strategies and the application of various machine learning algorithms. The basic dataset used for analysis was obtained from the ChEMBL206 database managed by the European Bioinformatics Institute.

The study follows a three-stage analysis process:

2.1. Molecular dataset preparation and Lipinski analysis

The activity classification of ERα inhibitors obtained from the ChEMBL206 database was performed based on experimental IC50 values. The IC50 value is a quantitative measure expressing the concentration required for an inhibitor to reduce the activity of its target protein by 50%. Molecules were classified into three categories according to their IC50 values: molecules with IC50 1000 nM were designated as 'active', molecules with IC50 10000 nM were designated as 'inactive', and molecules between these two values were designated as 'intermediate' (Krippendorff et al., 2007).

The drug-likeness properties of molecules were evaluated by calculating Lipinski's Rule of Five (Lipinski et al., 2001). According to this rule, a molecule is likely to be an orally active drug if it violates no more than one of the following criteria:

- Molecular weight (MW) - less than 500 Daltons
- Logarithm of octanol-water partition coefficient (LogP) - less than 5
- Number of hydrogen bond donors – less than 5 or fewer
- Number of hydrogen bond acceptors – less than 10 or fewer

Chemical space analysis was performed using these calculated Lipinski rules and activity classes, and MW-LogP distributions of molecules were visualized (Wagener & Van Geerestein, 2000).

2.2. Molecular descriptor calculation and structural characterization

Comprehensive molecular characterization was performed using the PADEL program. A total of 3153 different molecular descriptors were calculated for 3053 molecules. These descriptors provide a detailed characterization of the structural properties of molecules (atom and bond properties), topological characteristics (molecular connectivity), electronic properties (charge distribution), 2D and 3D structural properties, and physicochemical parameters. This comprehensive analysis also enabled the identification of molecular fingerprints. The obtained PADEL descriptors were combined with activity classes previously obtained from the ChEMBL206 database and determined according to IC50 values. During data preprocessing, molecules classified as the 'intermediate' class were excluded, and the task was reformulated as a binary classification problem including only active and inactive categories.

2.3. Machine learning and feature importance analysis

First, the prepared dataset was divided into 80% training and 20% validation sets using random sampling method while maintaining the proportions of active and inactive molecules. The 20% portion was reserved as an external validation set to evaluate the final model's performance. On the training set, a cross-validation method, which divides the dataset into different training-test subsets, was applied to control whether the models overfit the dataset and to evaluate generalization ability. Various algorithms based on ensemble learning approaches, which can make stronger predictions by combining multiple base learners, were used in the construction of classification models (Al-Thanoon et al., 2019).

Next, the Random Forest algorithm creates multiple decision trees by selecting random samples from the dataset using the bootstrap sampling method and determines the best split by using randomly selected features at each node (Breiman, 2001). The Bagging (Bootstrap Aggregating) algorithm, which works similarly, reduces the variance of the model by combining the predictions of decision trees trained in parallel on different subsets created by bootstrap sampling from the dataset (Breiman, 1996).

Then, XGBoost (eXtreme Gradient Boosting) and LightGBM (Light Gradient Boosting Machine) algorithms are advanced ensemble learning methods that work with the gradient boosting principle. These algorithms sequentially train weak learners that focus on correcting the errors of previous models, minimizing the gradient of the loss function at each step (Chen &

Guestrin, 2016; Ke et al., 2017). LightGBM is an optimized version that works faster than XGBoost, especially on large datasets.

Additionally, the ExtraTrees (Extremely Randomized Trees) algorithm, unlike standard decision trees, uses completely randomly selected threshold values in node splits. This approach reduces the variance of the model while increasing computational efficiency (Geurts et al., 2006). All these algorithms stand out with their ability to model complex relationships in high-dimensional data containing many variables.

Then, the classification performance of the created models was measured using various evaluation metrics. Specifically, accuracy (the ratio of correctly classified samples to old samples), sensitivity (the ratio of correctly predicted active molecules to all active molecules), and specificity (the ratio of correctly predicted inactive molecules to all inactive molecules) were calculated.

Moreover, ROC (Receiver Operating Characteristic) curves were drawn to evaluate the performance of the models at different decision threshold values, and the area under these curves (AUC - Area Under the ROC Curve) was calculated to measure overall classification success. The ROC curve is a graphical representation of sensitivity (true positive rate) and 1-specificity (false positive rate) across different threshold values. The AUC provides an overall measure of the model's ability to distinguish between classes; values close to 1 indicate perfect classification performance, whereas a value of 0,5 corresponds to random prediction (Sokolova et al., 2006; Byvatov et al., 2003).

Finally, a comprehensive feature importance analysis was performed through the trained models, and the molecular properties that contributed most to ERα inhibitor activity were determined. Feature importance ranking was obtained using the intrinsic feature importance mechanisms of ensemble learning models (especially Random Forest and XGBoost). Feature importance analysis makes important contributions to understanding structure-activity relationships and the rational design of new inhibitors.

3. Results

All analyses and model development studies used for activity prediction of ERα inhibitors were conducted in the Python programming language. The pandas and numpy libraries were utilized for data analysis and processing, while matplotlib and seaborn libraries were employed for visualization. For the development of machine learning models, scikit-learn, XGBoost, and LightGBM libraries were utilized. Specifically, the sklearn metrics module was used for model performance evaluation, and the internal mechanisms of ensemble learning models were employed for feature importance analyses. The findings of the study were examined under three main headings:

3.1. Drug-likeness analysis

The chemical space analysis of ERα inhibitors revealed important structure-activity relationships by visualizing the distribution of Molecular Weight (MW) and LogP values of molecules, as shown in Figure 1. In the chemical space plot, each point represents a molecule, where colors indicate activity classes (green: active, red: inactive, yellow: intermediate) and sizes represent pIC50 values (4: low activity, 6: moderate activity, 8: high activity, 10: very high activity).

The analysis results showed that active ERα inhibitors are generally concentrated within the molecular weight range of 200-1200 Da and LogP values between 2-10. This distribution is generally consistent with Lipinski's rule of five for drug-likeness and suggests that the molecules exhibit suitable properties for oral bioavailability. Particularly, the tendency of active molecules (green dots) to have higher MW and LogP values indicates that these physicochemical properties may be important determinants for ERα inhibitory activity.

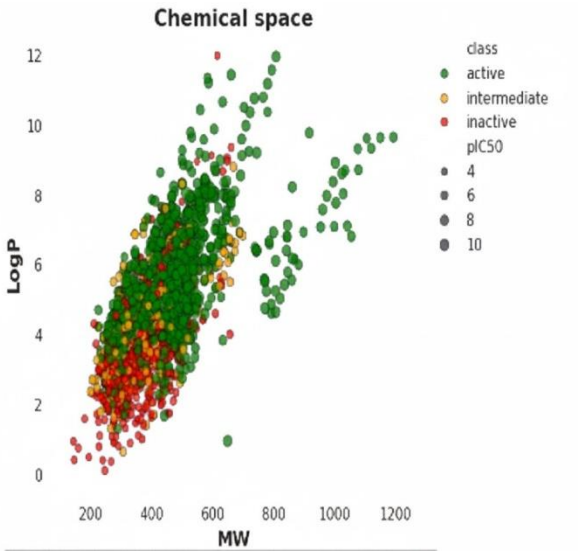


Figure 1. Chemical space distribution of ERα inhibitors

The clustering of inactive molecules (red dots) in the low MW and LogP region may result from these molecules' inability to form sufficient interaction surfaces with the target protein. Molecules with 'intermediate' activity (yellow dots) are generally located in the transition zone between active and inactive regions. This chemical space distribution provided valuable information for novel inhibitor design by enabling a better understanding of the structural features that determine the activity of ERα inhibitors, particularly emphasizing that MW and LogP values need to be within specific ranges for optimal activity.

3.2. Performance evaluation of classification models

In the machine learning approach applied for predicting the activity of ERα inhibitors, a systematic methodology was followed throughout the data preparation and model development processes. Initially, molecules obtained from the ChEMBL206 database were classified into three different activity classes based on their IC50 values, as presented in Table 1.

Table 1. Distribution of molecules by activity classes in the initial dataset

Activity Class	Number of Molecules	Percentage (%)
Active	1773	58.07
Inactive	725	23.75
Intermediate	555	18.18

In this classification, more than half of the total 2053 molecules were categorized as active, approximately one-fourth were determined as inactive, and the remaining small portion was classified in the 'intermediate' category. To optimize the prediction performance of classification models, molecules with intermediate activity values were removed from the dataset, and the problem was addressed using a binary classification approach as shown in Table 2.

Table 2. Distribution of the reorganized dataset for binary classification

Activity Class	Number of Molecules	Percentage (%)
Active	1773	70.98
Inactive	725	29.02

In the dataset organized for binary classification, the distribution shows approximately two-thirds active molecules and one-third inactive molecules. This distribution was maintained to ensure balanced class representation during model training and validation. The dataset was divided into 80% training and 20% validation sets for model development and independent evaluation of the developed models, while ensuring the preservation of the original class distributions as detailed in Table 3. This strategic division enabled the models to make reliable predictions for both classes.

Table 3. Class distributions of training and validation sets

Set	Class	Number of Molecules	(%)
Training	Active	1418	70.97
	Inactive	580	29.03
Validation	Active	355	71.00
	Inactive	145	29.00

To assess the generalization performance of the models, the training set was subjected to 5-fold cross-validation, and performance metrics were derived from the external validation set. During the model development process, various machine learning algorithms were evaluated as summarized in Table 4. Among ensemble learning algorithms, XGBoost and LightGBM stood out with the highest accuracy rates. This result demonstrates that gradient boosting approaches can successfully capture patterns in complex molecular data. Other ensemble models including ExtraTrees and Random Forest, along with a basic classifier, Logistic Regression, also exhibited similarly high performance. It was observed that all models, particularly ensemble-based algorithms, could effectively distinguish between active and inactive molecules.

Table 4. Performance scores of the machine learning models

Model	Accuracy	Sensitivity	Specificity	F1-Score
XGBoost	94.80	90.0	97.0	0.91
LightGBM	94.60	90.0	97.0	0.91
ExtraTrees	94.00	90.0	95.0	0.90
Random Forest	93.80	89.0	96.0	0.89
Logistic Regression	93.40	88.0	96.0	0.89
Bagging	91.60	79.0	97.0	0.84

To comprehensively evaluate the classification performance of the models, ROC (Receiver Operating Characteristic) curves were generated, and the areas under the curves (AUC) were calculated as illustrated in Figure 2. ROC analysis results demonstrate that all models exhibit strong discriminative performance.

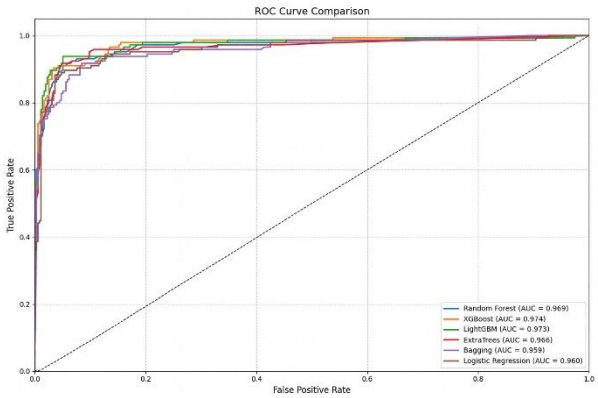
Particularly, gradient boosting-based models (XGBoost and LightGBM) achieved the highest classification success. This was followed by other ensemble-learning algorithms (Random Forest and ExtraTrees), along with Logistic Regression and Bagging algorithms. When examining F1-scores, it is also observed that ensemble learning models exhibit balanced precision and sensitivity. The distinct separation of ROC curves from the random prediction line (diagonal) confirms that all developed models are highly successful in molecular activity prediction and can produce stable and reliable results even at different decision thresholds.

3.3. Feature importance analysis in molecular activity prediction

In addition to performance evaluation of the developed models, feature importance analysis was conducted to understand which molecular descriptors are more determinant in predicting estrogen receptor alpha activity, as demonstrated in Figure 3. This analysis visualizes the features most utilized by each machine

learning model during the prediction process and their relative importance degrees.

Figure 2. ROC curve comparison of different machine learning models



3.3.1 Graphical analysis of feature importance

The feature importance graph presented in Figure 3 illustrates the top 20 molecular descriptors of the three best-performing machine learning models (XGBoost, LightGBM, and ExtraTrees). For each model, the descriptors are ranked in descending order based on their importance scores. A notable observation from the graph is that each model employs a distinct feature, an importance scoring scale. For instance, the feature scores in the LightGBM model range from 0 to 80, whereas those in the XGBoost model range from 0 to 0.05. This variation arises from the internal feature evaluation mechanisms specific to each model.

For the sake of clarity and brevity, graphical representations are provided in the main text only for the three aforementioned models (XGBoost, LightGBM, and ExtraTrees). Feature importance results for the remaining models are available in the Supplementary Materials.

3.3.2 Prominent descriptors by models

Table 5 summarizes the groups that the most frequently prominent molecular descriptors belong to and the chemical properties they represent.

This feature importance analysis demonstrates that specific molecular descriptors and structural features are more effective in predicting estrogen receptor alpha activity. Particularly, the prominence of Pubchem fingerprint descriptors and atom pair descriptors in most models suggests that the molecular properties represented by these descriptors are critical for bioactivity prediction.

The important descriptors that stand out in different models are as follows: In the Random Forest model, PubchemFP308, PubchemFP667, and SubFP169; in the XGBoost model, PubchemFP667, PubchemFP527, and PubchemFP597; in the LightGBM model, APC2D3_C_C, APC2D5_C_C, and APC2D6_C_C atom pair descriptors; in the ExtraTrees model, KRFP1148, PubchemFP667, and KRFP2949; in the Bagging model, PubchemFP667, PubchemFP527, and KRFPC3013; and in the Logistic Regression model, KRFPC348, KRFPC607, and FP834 are observed as the most determinative features.

When examining the most frequently repeated features across all models, PubchemFP667 appears among important descriptors in five different models (Random Forest, XGBoost, LightGBM, ExtraTrees, and Bagging). Similarly, PubchemFP527 was also found to be commonly important in XGBoost, LightGBM, and Bagging models.

Figure 3. Feature importance analysis of the models

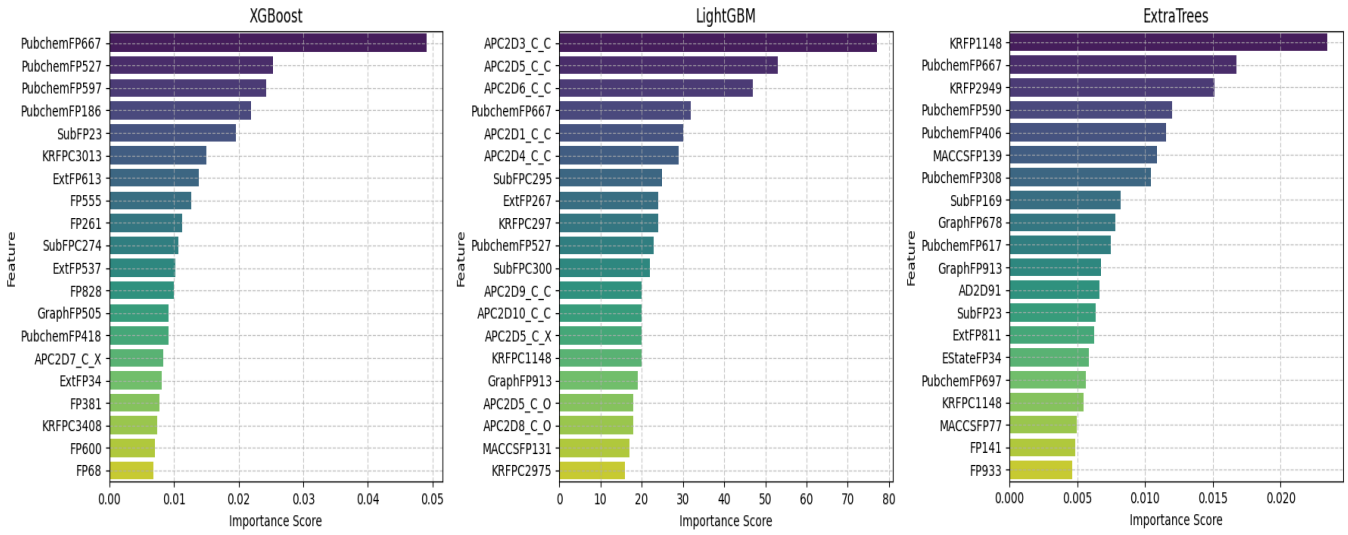


Table 5. Prominent molecular descriptor groups and their chemical meanings

Descriptor	Example Descriptors	Chemical Meaning
PubchemFP	PubchemFP667, PubchemFP527	Pubchem fingerprint. Represents specific structural features and functional groups.
KRFP	KRFP1148, KRFP2949	Klekota-Roth fingerprint. Represents structural motifs associated with bioactivity.
SubFP	SubFP169, SubFP23	Substructure fingerprint. Represents ring systems and functional groups.
APC2D	APC2D3_C_C, APC2D5_C_C	Atom pair count. Represents distances between specific atom pairs.
EStateFP	EStateFP34	Electrostatic fingerprint. Represents electronic distribution properties.
GraphFP	GraphFP678, GraphFP913	Graph-based fingerprint. Represents topological properties of the molecule.
MACCSFP	MACCSFP139, MACCSFP131	MACCS keys. Represents standard structural fragments.
ExtFP	ExtFP613, ExtFP267	Extended fingerprint. Represents extended molecular properties.

4. Discussion

In this study, the effectiveness of statistical learning approaches in predicting the activity of ERα inhibitors was comprehensively evaluated. Analyses conducted on molecular data obtained from the ChEMBL database for the ChEMBL206 target reveal the potential value of in-silico methods in the drug discovery process.

The results obtained in the study demonstrate the success of ensemble learning algorithms, particularly in molecular activity prediction.

The chemical space analysis conducted in the first phase of the research showed that active ERα inhibitors cluster around specific physicochemical properties. The tendency of molecules with molecular weights between 200-1200 Da and LogP values between 2-10 to show high activity is an important finding that should be considered in the design of new inhibitors. This distribution is consistent with Lipinski's drug-likeness rules and suggests that the developed molecules may also exhibit suitable properties in terms of oral bioavailability.

In the performance evaluation of machine learning models, the XGBoost algorithm showed the highest success with 94.8% accuracy. This was followed by LightGBM with 94.6% accuracy, ExtraTrees with 94.0%, Random Forest with 93.8%, and Logistic Regression algorithms with 93.4%. These results reveal that gradient boosting-based models (XGBoost and LightGBM) are the most suitable algorithms for predicting estrogen receptor activity. The high accuracy of the XGBoost model demonstrates that the algorithm can successfully learn complex molecular feature-activity relationships.

When examining the sensitivity and specificity values of the models, XGBoost and LightGBM algorithms exhibited the most balanced performance with 90% sensitivity and 97% specificity. This shows that the models can predict both active and inactive molecules with high accuracy. Particularly when examining F1-score values, gradient boosting-based algorithms show superior performance with a value of 0.91. These results prove that the developed models can make balanced and reliable predictions for both classes.

ROC curve analysis results show that all models exhibit strong discriminative performance. XGBoost achieved the highest classification success with an AUC value of 0.974. This was followed by LightGBM (0.973 AUC), Random Forest (0.969 AUC), and ExtraTrees (0.966 AUC) algorithms. The distinct separation of ROC curves from the random prediction line confirms that the developed models can produce stable and reliable predictions even at different decision thresholds.

One of the most notable findings of the study is the identification of the importance of specific descriptors in molecular activity prediction. Feature importance analysis enabled the identification of common descriptors that stand out in different models. The fact that the PubchemFP667 descriptor appears among important features in five different models (XGBoost, LightGBM, Random Forest, ExtraTrees, and Bagging) shows that this structural feature plays a critical role in ERα inhibition. Similarly, PubchemFP527 was also found to be commonly important in XGBoost, LightGBM, and Bagging models.

The APC2D atom pair descriptors (APC2D3_C_C, APC2D5_C_C, and APC2D6_C_C) that stand out in the LightGBM model emphasize the importance of specific distances between carbon-carbon atoms in predicting molecular activity. These findings suggest that specific atom-atom interactions may have a direct effect on activity for ERα inhibitors. The high importance scores of KRFP descriptors (especially KRFP1148 and KRFP2949) in the ExtraTrees model indicate that Klekota-Roth fingerprint descriptors can successfully capture structural motifs associated with bioactivity.

When compared with similar studies reported in the literature, the results obtained are quite promising. In a

comprehensive analysis conducted by DiMasi and colleagues, it was reported that the total cost of the traditional drug discovery process, including post-approval R&D costs, reaches $2.87 billion, and these costs increase at an annual rate of 8.5% (DiMasi et al., 2016). Considering this high cost and time burden, the developed models will enable rapid and reliable evaluation of potential ERα inhibitors. The results of feature importance analysis provide important clues for designing new ERα inhibitors. The prominence of Pubchem fingerprint descriptors and atom pair descriptors in most models shows that the molecular properties represented by these descriptors are critical for bioactivity prediction. This information can be guided in determining which structural features should be prioritized in the drug design process. For example, focusing on molecules carrying the structural features represented by PubchemFP667 could be an important strategy in designing candidate molecules with high activity potential.

The results of this study clearly reveal the potential value of in-silico methods in the drug discovery process. The developed models not only provide high accuracy in predicting the activity of ERα inhibitors but also offer valuable insights for designing new inhibitors. The combination of machine learning algorithms and feature importance analysis creates a powerful framework for rational drug design. This approach can be used as a valuable tool in filtering and prioritizing candidate molecules before proceeding to laboratory tests, thus significantly reducing the cost and duration of the traditional drug discovery process.

5. Conclusions

This study comprehensively demonstrated the power of machine learning-based QSAR models in predicting the inhibitory activity of estrogen receptor alpha (ERα) inhibitors. Using molecular fingerprints and atom pair descriptors, high-performance models such as XGBoost and LightGBM achieved accuracy rates exceeding 94% and AUC values above 0.97, confirming the robustness of ensemble learning approaches in molecular classification tasks. The balanced sensitivity and specificity values, along with superior F1-scores, indicate the models' capability to make reliable predictions across both active and inactive classes.

Importantly, feature importance analyses revealed critical structural patterns influencing bioactivity. Descriptors such as PubchemFP667 and APC2D3_C_C emerged as consistently influential across several models, highlighting their role in ERα inhibition. These results suggest that specific substructural features and atom-level interactions significantly contribute to biological activity, offering valuable guidance for rational drug design.

Considering the high cost and time constraints of traditional drug discovery, the developed in silico framework presents a rapid, reliable, and cost-effective alternative for the early-stage screening and prioritization of candidate molecules. Overall, the integration of statistical learning techniques with chemical informatics holds strong potential for accelerating the identification and optimization of novel ERα inhibitors, contributing meaningfully to the discovery of new therapeutic options for hormone-related cancers such as breast cancer.

References

Al-Thanoon, N. A., Qasim, O. S., Algamal, Z. Y., 2019. A new hybrid firefly algorithm and particle swarm optimization for tuning parameter estimation in penalized support vector machine with application in chemometrics. Chemometrics and Intelligent Laboratory Systems, 184, 142–152.

Ali, S., Coombes, R. C., 2002. Endocrine-responsive breast cancer and strategies for combating resistance. Nature Reviews Cancer, 2(2), 101–112.

Anderson, E., 2002. The role of oestrogen and progesterone receptors in human mammary development and tumorigenesis. Breast Cancer Research, 4(5), 197–201.

Arciniegas, F., Bennett, K., Breneman, C., Embrechts, M. J., 2000. Molecular database mining using self-organizing maps for the design of novel pharmaceuticals. In Intelligent Engineering Systems through Artificial Neural Networks: Smart Engineering System Design, 10, 477–481.

Breiman, L., 1996. Bagging predictors. Machine Learning, 24(2), 123–140.

Breiman, L., 2001. Random forests. Machine Learning, 45(1), 5–32.

Byvatov, E., Fechner, U., Sadowski, J., Schneider, G., 2003. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. Journal of Chemical Information and Computer Sciences, 43(6), 1082–1089.

Carracedo-Reboredo, P., Liñares-Blanco, J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F. J., Carballal, A., Maojo, V., Pazos, A., Fernandez-Lozano, C., 2021. A review on machine learning approaches and trends in drug discovery. Computational and Structural Biotechnology Journal, 19, 4538-4558.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Deroo, B. J., Korach, K. S., 2006. Estrogen receptors and human disease. Journal of Clinical Investigation, 116(3), 561–570.

DiMasi, J. A., Grabowski, H. G., Hansen, R. W., 2016. Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of Health Economics, 47, 20-33.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Leach, A. R., 2017. The ChEMBL database in 2017. Nucleic Acids Research, 45(D1), D945–D954.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine Learning, 63(1), 3–42.

Gümüştaş, E. and Çakmak Pehlivanlı, A.Ç., 2021. In-silico mutajenisite tahmininde istatistiksel öğrenme modeli.

Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 25(2), 365–370.

Huang, P., Chandra, V., Rastinejad, F., 2010. Structural overview of the nuclear receptor superfamily: insights into physiology and therapeutics. Annual Review of Physiology, 72, 247–272.

Jensen, E. V., Jordan, V. C., 2003. The estrogen receptor: a model for molecular medicine. Clinical Cancer Research, 9(6), 1980–1989.

Jordan, V. C., 2003. Targeting anti-hormone resistance in breast cancer: a simple solution. Annals of Oncology, 14(7), 969–970.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T. Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems (NIPS 2017), 30, 3146-3154.

Krippendorff, B. F., Lienau, P., Reichel, A., Huisinga, W., 2007. Optimizing classification of drug-drug interaction potential for CYP450 isoenzyme inhibition assays in early drug discovery. Journal of Biomolecular Screening, 12(1), 92–99.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature, 521, 436-444.

Lipinski, C. A., 2004. Lead- and drug-like compounds: the rule-of-five revolution. Drug Discovery Today: Technologies, 1(4), 337-341.

Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews, 46(1-3), 3–26.

Musgrove, E. A., Sutherland, R. L., 2009. Biological determinants of endocrine resistance in breast cancer. Nature Reviews Cancer, 9(9), 631–643.

Qasim, O. S., Algamal, Z. Y., 2018. Feature selection using particle swarm optimization-based logistic regression model. Chemometrics and Intelligent Laboratory Systems, 182, 41-46.

Sakri, S. B., Abdul Rashid, N. B., Muhammad Zain, Z., 2018. Particle swarm optimization feature selection for breast cancer recurrence prediction. IEEE Access, 6, 29637-29647.

Sebaugh, J. L., 2011. Guidelines for accurate EC50/IC50 estimation. Pharmaceutical Statistics, 10(2), 128–134.

Shaker, B., Ahmad, S., Lee, J., Jung, C., Na, D., 2021. In silico methods and tools for drug discovery. Computational Biology in Medicine, 137, 104851.

Sliwoski, G., Kothiwale, S., Meiler, J., Lowe, E. W., 2013. Computational methods in drug discovery. Pharmacological Reviews, 66(1), 334–395.

Sokolova, M., Japkowicz, N. & Szpakowicz, S., 2006. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. In: A. Sattar & B. Kang, eds., AI 2006: Advances in Artificial Intelligence, Lecture Notes in Computer Science, vol. 4304, Springer, Berlin Heidelberg, 1015–1021.

Wagener, M., Van Geerestein, V. J., 2000. Potential drug and non-drugs: prediction and identification of important structural features. Journal of Chemical Information and Computer Sciences, 40(2), 280–292.

Yap, C. W., 2011. PaDEL-descriptor: An open-source software to calculate molecular descriptors and fingerprints. Journal of Computational Chemistry, 32(7), 1466–1474.

Zhang, J. H., Chung, T. D., Oldenburg, K. R., 2000. Confirmation of primary active substances from high throughput screening of chemical and biological populations: a statistical approach and practical considerations. Journal of Combinatorial Chemistry, 2(3), 258–265.