Predicting active compounds for lung cancer based on quantitative structure-activity relationships

Hamza Hanafi[1], Badr Dine Rossi Hassani[2], M'hamed Aït Kbir[1]
[1]Intelligent Automation Team, STI Doctoral Center, Abdelmalek Essaadi University, Tangier, Morocco
[2]LABIPHABE Laboratory, STI Doctoral Center, Abdelmalek Essaadi University, Tangier, Morocco

Article Info

Article history:

Received Jan 18, 2023
Revised Apr 7, 2023
Accepted Apr 15, 2023

Keywords:

Bioinformatics
Diagnosis of cancer
Drug discovery
Drug target
Lung cancer

ABSTRACT

Recently, advancements in computational and artificial intelligence (AI) methods have contributed in improving research results in the field of drug discovery. In fact, machine learning techniques have proven to be especially effective in this regard, aiding in the development of new drug variants and enabling more precise targeting of specific disease mechanisms. In this paper, we propose to use a quantitative structure-activity relationship-based approach for predicting active compounds related to non-small cell lung cancer. Our approach uses a neural network classifier that learns from sequential structures and chemical properties of molecules, as well as a gradient boosting tree classifier to conduct comparative analysis. To evaluate the contribution of each feature, we employ Shapley additive explanations (SHAP) summary plots to perform features selection. Our approach involves a dataset of active and non-active molecules collected from ChEMBL database. Our results show the effectiveness of the proposed approach when it comes to predicting accurately active compounds for lung cancer. Furthermore, our comparative analysis reveals important chemical structures that contribute to the effectiveness of the compounds. Thus, the proposed approach can greatly enhance the drug discovery pipeline and may lead to the development of new and effective treatments for lung cancer.

Corresponding Author:

Hamza Hanafi
Intelligent Automation Team, STI Doctoral Studies Center, Abdelmalek Essaadi University
Tangier, Morocco
Email: hamzahanafi1@gmail.com/hamza.hanafi@etu.uae.ac.ma

1. INTRODUCTION

Research works conducted in the field of drug discovery are important and contribute to the improvement of healthcare quality. Developing a new drug is a long and complex process that relies on the translation of a new molecular target into a proven therapy with efficient results. Drug discovery is one of the most outstanding scientific tasks. Advances in computational biology have broadly improved drug discovery pipelines. Classical methods directed towards this goal are time-consuming and expensive [1]. Therapeutic studies are crucial for designing new drugs for the benefit of patients, as well as for public health reasons [2].

Nowadays, computational techniques have expanded their focus and greatly improved pipelines in the field of pharmacological medicine, as they have demonstrated successful results compared to traditional methods. Moreover, the remarkable amount of biological data publicly available and carefully stored in repositories has enabled researchers to explore numerous computational-based methodologies. Predictive modeling is one of the most widely applied techniques to enhance drug discovery pipelines. Machine learning (ML) techniques can be utilized to construct models that effectively classify drugs into relevant therapeutic categories and accurately detect and classify various stages of tumors [3], [4]. Additionally, ML methods can be used to design new drugs based on the chemical properties of studied molecules [5].

Quantitative structure-activity relationship (QSAR) methods are techniques that apply ML in order to learn from the relationships among the chemical structure and the biological activity [6] of molecules, additionally, they have helped to establish an empirical statistical model for the computational chemistry toolkit [7]. The chemical structure of molecules is subject of calculations of molecular descriptors that describe essentially the physical and chemical properties that distinguish one molecule from another. QSAR-based models can provide insights on which chemical properties are important to inhibit a biological process. Such information will be of great interest to biologists and chemists in their design of future molecules in order to have more robust properties.

Bioinformatic methods have successfully enabled researchers to study molecules from a system level perspective. It uses computational processes to integrate knowledge and expertise from genomics, proteomics, transcriptomics, population genetics, and molecular phylogenetics. Bioinformatic analysis has enhanced drug target identification and drug candidate screening. Moreover, it facilitates predictions of drug resistance, minimized side effects, and has become more essential in drug discovery [8]. Thus, numerous ML-based algorithms have been proposed to predict interactions among biological entities, as well as to design new drugs with similar properties for specific medical treatments. One of the main challenges to build an efficient ML classifier is the absence of good quality data. In fact, the available biological data is heterogeneous and requires a preprocessing step before initiating the training process of the ML models. Moreover, in cancer classification problems, most of the available datasets are imbalanced; as there are extensively more non-active molecules than active ones [9].

Our contribution is to build a classifier able to predict active compounds for lung cancer. We inferred active and non-active molecules from ChEMBL database to constitute our dataset. We computed fingerprints descriptors of collected molecules to learn from them. We took full advantage of the chemical characteristics and the structure of the molecules to build a sequential neural network model. Furthermore, we conducted a comparative study between the Multilayer perceptron neural network and the gradient boosting tree classifiers to analyze features contribution for each model in order to identify important chemical structures of active molecules for lung cancer.

This paper is a part of a series of research carried out by a team of our laboratory, interested to exploring biological data using datamining machine learning tools [10]–[12]. The paper is organized as: in section 2 we present some related works, our approach is presented in section 3. The obtained results are discussed in section 4 and finally the conclusion.

2. RELATED WORK

Experiments to identify new therapeutic targets are aimed at investigating novel molecules and improve bioavailability of drugs. Traditional methods applied in drug discovery rely on the physical and chemical structure of the studied molecules. Genome-wide association studies (GWAS) screen a large number of genomes to identify associations between genetic variants and non-disease traits [13]. This approach has been widely used to identify single nucleotide polymorphisms (SNPs) associated with diseases and greatly improved our understanding of biological processes [14]. Identification of drug target sites is another methodology that many studies rely on. It refers to the discovery of interactions among diverse compounds and protein targets in the human body. Lee *et al.* [15] experimentally demonstrated that the duration of in vivo drug-target binding is highly affected by the drug-target resistance. In another study on targeted therapies for lung cancer predictions, Larsen *et al.* [16] suggested to integrate genome-wide tumor analysis along with drug-targeted responsive phenotypes to investigate new therapeutic strategies. This approach requires further knowledge on the binding sites. Moreover, it involves prior knowledge of related pathways to develop effective targeted therapeutics.

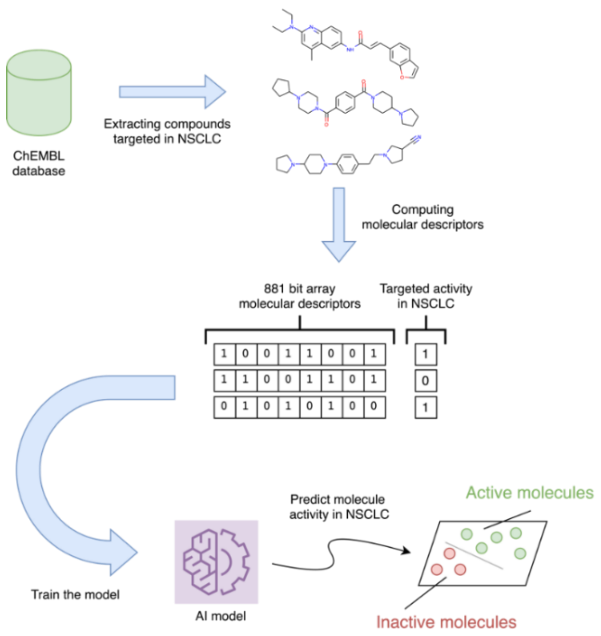
Structure-based approaches have significantly enhanced virtual screening, de novo design, and lead optimization [17], [18], based on the availability of ligand structures. On this subject, Almeida *et al.* [19] used multiple ligand-based virtual screening approaches to investigate novel potential MARK-3 Inhibitors in cancer. Similarly, Li *et al.* [20] suggested a nanoparticle-mediated targeted drug as a novel therapeutic for hepatocellular carcinoma using ligands that recognize hepatoma cells. The main disadvantage of this approach is that it cannot be used in situations where ligands are unknown. Similarity-based methods have also been used to design novel compounds. QSAR is a methodology that suggests structurally similar compounds tend to possess similar biological activities [21]. Numerous studies based on this approach calculate a similarity score among drug profiles to discover potential drug-drug interactions. Vilar and Hripcsak [22] used several drug profiles to compute a similarity score between multiple compounds. Correspondingly, Ferdousi *et al.* [23] compared diverse molecular profiles and found that the structural profile is the most optimal metric to predict drug-drug interactions. The major disadvantage of this method is the choice of a suitable threshold for the computed similarity; which is highly affected by the quality of the used dataset and false positive interactions. Likewise, classical methods used in drug discovery are time consuming. Besides, they are less accurate because of the number of reported ADRs.

Computational methods have significantly changed the way novel drugs are designed. Drugs discovery pipelines have been largely enhanced and improved our understanding of biological processes. Biological networks are a great way to represent chemical interactions as they have helped to integrate and create a model of diverse heterogeneous biological data. Hanaf *et al.* [12], proposed a network-based method combined with an ML algorithm to classify and predict interactions between genes, drugs, and diseases. Similarly, they were able to rank the top 20 gene-drug pairs related to lung cancer. Topological data analysis has recently been used to study large-scale biological data. Hanafi *et al.* [24], built a biological network using data integration methods and explored numerous graph properties to evaluate potential gene-disease interactions. Huang *et al.* [25] about drug repositioning for non-small cell lung cancer (NSCLC), the authors combined topological parameter-based classification and ML algorithms to explore potential therapeutic drugs for NSCLC, they successfully suggested promising drugs for treating early and late-stage lung cancer that were supported by the literature and appeared highly effective in clinical trials and in vitro. Similarly, in a study about identification of small potent molecule inhibitors to target Src kinase as a therapeutic strategy for lung cancer Weng *et al.* [26], constructed a computational model for the *in silico* screening of Src inhibitors and evaluated the effect of potential candidate compounds based on a QSAR model. The obtained results were promising, as the candidate compounds used revealed a significant inhibitory effect against Src activity.

In this paper, we present a computational QSAR-based model, combined with a tree-based classifier and a neural network model, to predict novel targeted compounds in lung cancer. We created a dataset of compounds related to NSCLC from the ChEMBL database. We computed molecular descriptors of the molecules as an 881-bit array, which we used as input features for the learning tasks. Furthermore, to evaluate our models, we conducted a feature engineering step and compared feature contributions using the SHAP values method.

3. OUR APPROACH

Our study follows a very meticulous approach to propose active compounds for lung cancer. The overall methodology is described in Figure 1. We started by collecting bioactive compounds related to non-small cell lung cancer from the ChEMBL database to construct our dataset. Afterwards, we clustered the compounds into two groups: highly active and non-active drugs, based on their inhibition concentration value at 50%, denoted as IC50. The lower the IC50, the more likely the drug is effective in inhibiting NSCLC. Then, we computed molecular descriptors and initiated two learning tasks to build our models and learn from the chemical characteristics of the calculated molecular descriptors.


ChEMBL database
Extracting compounds targeted in NSCLC
Computing molecular descriptors
881 bit array molecular descriptors
Targeted activity in NSCLC
Train the model
AI model
Predict molecule activity in NSCLC
Active molecules
Inactive molecules

1	0	0	1	1	0	0	1	1
1	1	0	0	1	1	0	1	0
0	1	0	1	0	1	0	0	1

Figure 1. Overall approach followed to predict compounds activity for non-small cell lung cancer

3.1. Dataset construction

We constructed our dataset from the ChEMBL database; it is a discovery platform that covers drug-like compounds [27], [28]. It offers a large variety of data related to drugs and provides insights, tools, and resources for drug discovery. Researchers use ChEMBL to make associations between diseases and their relevant targets. It also helps identify small molecules that can be used to target newly sequenced genomes. ChEMBL integrates its content primarily from the scientific literature, making it a great and accurate tool for *in silico* drug design. The collected compounds have two features to predict their binding activity. Table 1 shows some samples from our dataset.

The IC50 measures the potency of a molecule in inhibiting a biological process by 50%. It indicates how much of a substance is needed to inhibit half of a given process. Consequently, drugs with lower IC50 values are highly active and have a value of 1 in the column “Activity in NSCLC”. Conversely, drugs with higher IC50 values are less active and have a value of 0 in the column “Activity in NSCLC”.

Table 1. Some samples from the used dataset

Compound identifier in ChEMBL	IC50 value (nM)	Activity in NSCLC
CHEMBL133389	250	1
CHEMBL336075	10,000	0
CHEMBL130758	750	1
CHEMBL372987	7,000	0
CHEMBL201490	18	1
CHEMBL131038	10,000	0
CHEMBL185	8,400	0
CHEMBL159	5	1
CHEMBL11359	6,700	0
CHEMBL132876	40,300	0

3.2. Molecular descriptors

Molecular descriptors can be defined as a way to encode the chemical structure of molecules into numbers, typically represented as an array of bits. Each numerical value denotes the presence or absence of a certain pattern, such as a hydrogen bond, atom, or fragment. They are used to explore physicochemical and topological properties to establish the basis for *in silico* predictive QSAR-based models and are also useful in performing similarity searches in molecular libraries.

We used the PaDEL-descriptor [29] software to calculate the molecular descriptors of our collected molecules. The software was developed by the National University of Singapore with the aid of the Chemistry Development Kit. It uses the simplified molecular input line entry system (SMILES) to compute hundreds of molecular descriptors and fingerprints. SMILES is the simplest way to represent a molecule based on a line notation [30]. It is a way to encode a chemical structure using notations that can be read and understood by a computer. The ChEMBL database provides the SMILES notation for the collected compounds, and Table 2 shows some of our collected molecules with their corresponding SMILES notation.

PubChem is a chemistry database that covers substances, compounds, and bio-assays. It defines a binary substructure fingerprint for chemical structures. Each compound’s SMILES in our dataset was encoded into an array of 881 bits consisting of physicochemical properties defined by PubChem. Table 3 shows a summary description of the bits used by PubChem descriptors.

Table 2. Some samples of collected compounds and their SMILES notation

Compound identifier in ChEMBL	SMILES
CHEMBL133389	*O=C1C=CC(=O)c2c(O)c(Oc3ccccc3)c(Cl)c(O)c21*
CHEMBL336075	*Cc1ccc[n+](C2=C([O-])C(=O)c3c(O)ccc(O)c3C2=O)c1*
CHEMBL130758	*COc1c(Cl)c(Cl)c(OC)c2c1C(=O)C=CC2=O*
CHEMBL372987	*COC(=O)c1[nH]c2ccc(Cl)cc2c1Sc1ccccc1OC*
CHEMBL201490	*COC(=O)c1[nH]c2ccc(Cl)cc2c1Sc1cc(OC)cc(OC)c1*
CHEMBL131038	*COc1cc(OC)c2c3c(oc2c1)C(=O)c1c(OC)ccc(OC)c1C3=O*
CHEMBL185	*O=c1[nH]cc(F)c(=O)[nH]1*
CHEMBL132876	*O=C1c2ccccc2C(=O)c2c1oc1cc3c(cc21)OCO3*

3.3. Learning tasks

The dataset contains a collection of 142,852 molecules clustered into active and non-active groups based on the IC50 value. We allocated 90% of our data to build the training set and 10% for the test set. The target feature used to predict, activity in NSCLC, can take discrete values which are: 0 for non-active compounds and 1 for highly active compounds. We carried out a pre-processing step which consisted of reducing the number of features to use as an input to the ML model. The initial number of features is 881. Low variance features have been removed using a variance threshold of 0.15, which means dropping the feature where 85% of values are similar. We ended up with 175 features with high variance that will present a good set to allow the model to detect regularities present in the used dataset. A low number of features is also useful to perform optimal training phase experience.

To build our neural network, we used the Keras [31] library to define a multilayer perceptron model for binary classification. Physiochemical properties were fed into a sequential model, which consists of three hidden layers. Each layer is a dense class. The input layer size is 175 in order to be mapped to the feature vector. Then, the three hidden layers have 50, 10, and 2 neurons, respectively, with rectified linear unit (ReLU) as the activation function. The output layer has one node that uses the sigmoid activation function. We represented the physiochemical properties to the network with a single output value. We trained our model using binary cross-entropy as the loss function and the Adam method as the optimizer. Similarly, we applied a gradient boosting tree classifier to learn from a molecular descriptors array to predict the activity of compounds in NSCLC. We used the extreme gradient boosting (XGBoost) algorithm from the Scikit-Learn [32], [33] implementation to train and evaluate our model’s performance.

Table 3. Description of bits defined by PubChem

PubChem bit position range	Description
From 0 to 114	These bits check for the existence or count of individual chemical atoms
From 115 to 262	These bits check for the existence of rings
From 263 to 326	These bits check for the existence of bonded atom pairs, regardless of their count and order
From 327 to 448	These bits check for the existence of atom nearest neighbor patterns, taking into account aromaticity significant bonds
From 445 to 459	These bits check for the existence of detailed atom neighborhood patterns, regardless of count, but where bond orders are specific
From 460 to 712	These bits check for the existence of simple SMARTS patterns, regardless of count, but where bond orders are specific and bond aromaticity matches both single and double bonds
From 713 to 880	These bits check for the existence of complex SMARTS patterns, regardless of count, but where bond orders and bond aromaticity are specific

4. RESULTS AND DISCUSSION

Our neural network is a sequential feedforward model consisting of 3 hidden layers. The performance of the model was evaluated based on the log loss function as well as the reached accuracy during the training process over 100 epochs. The data was shuffled and split into portions called batches, with each batch consisting of 10 samples. During the learning process, the model loops over all these batches in each epoch and updates the model. Figure 2 shows the evolution over training cycles of the log loss. The model achieved an accuracy score of 0.96 with a log loss of 0.1166.

Similarly, we calculated the log loss of the decision tree-based classifier to evaluate its performance over 100 epochs. We performed tuning using the grid-search function to find optimal values for the hyperparameters of the model, which reached its highest performance with 100 boosted trees for a max depth of 6 levels. Figure 3 shows the obtained curve of the log loss function, which achieved a value of 0.008 for the validation set. The plot illustrates a decreasing curve in the log loss function. In addition, the model achieved an F1 score of 0.86 for both predicted classes.

This gives us a snapshot of the training process which is successful for the two models, the XGBoost model is more effective than the neural network-based model. We can see that both models can achieve highest prediction performances. However, the number of predictors (175) is still high. Consequently, finding relevant features is a crucial step to depict important structures of active molecules in lung cancer. Moreover, it will help set up appropriate model parameters that will enhance the classification results of our method when evaluated on unseen molecules. For that reason, we calculated the most relevant features to fully utilize the capabilities of our models, which can easily capture patterns within the structure of novel molecules and reduce the number of hyperparameters that need to be tuned. We plotted the shapely values using the Shapley additive explanations (SHAP) method [34], a way to explain how the model is estimating a prediction class for a given molecule. It is also a way to measure feature contribution and to find the most relevant features for the used dataset. Figures 4 and 5 depict the key features identified by the artificial neural network (ANN) and XGBoost models, respectively, highlighting the essential molecular patterns utilized by both models to make predictions. Remarkably, there are 11 common features that both models leverage, suggesting that these features play a critical role in distinguishing active and inactive drugs in NSCLC. To provide a comprehensive overview, Table 4 summarizes these 11 features, shedding light on the structural characteristics that underlie drug efficiency in NSCLC.

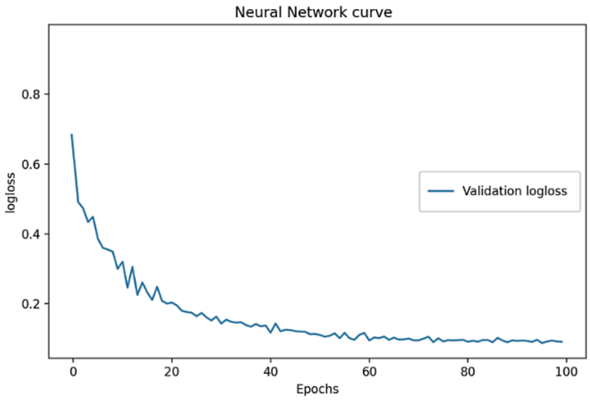


Figure 2. Log loss curve obtained for the artificial network model

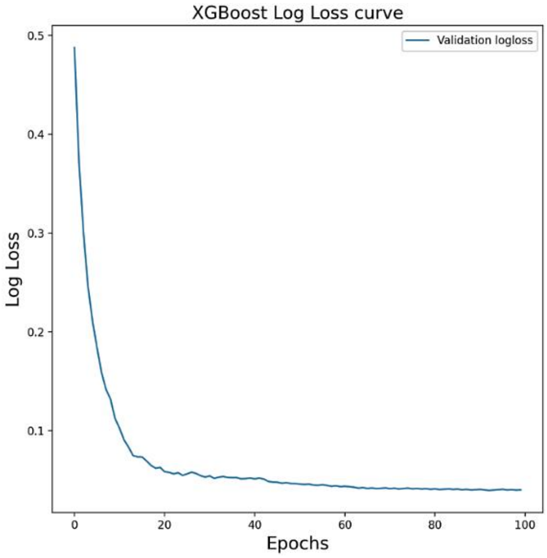


Figure 3. Log loss curve obtained for the XGBoost model

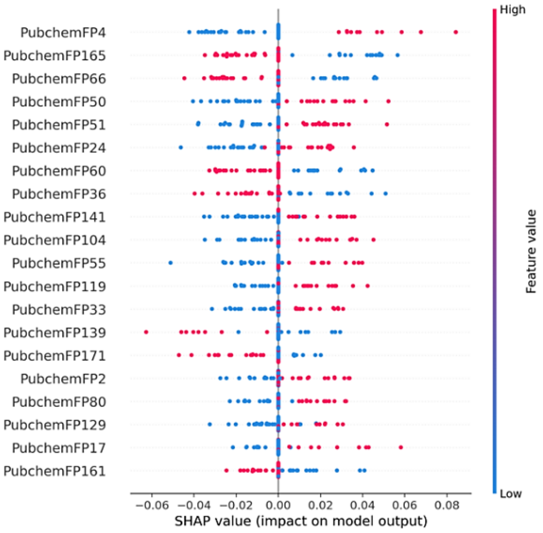


Figure 4. Feature contribution in the artificial neural network model

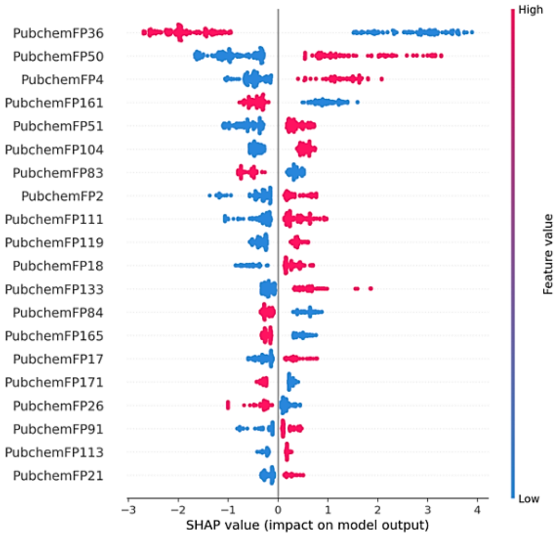


Figure 5. Feature contribution in the XGBoost model

Now, we have performed the training task for our XGBoost model using a reduced number of features. The 11 most relevant features were used as input for the learning process. The model achieved an F1 score of 0.98 with a final mean squared error of 0.02. This score is higher compared to that reached by the study in [35], where an F1 score of 0.76 was obtained. Moreover, we compared the ability of our model to predict drugs that were revealed as highly active in the study [35]. Then, we ranked top-10 highly active molecules in lung cancer, and Table 5 shows a list of these drugs.

The top-10 list of molecules we obtained were all supported by the literature. Erlotinib, which ranked 1, is an oral anticancer drug that inhibits the epidermal growth factor receptor responsible for excessive cell development in malignant lung tumors [36]. Paclitaxel, which ranked 2, is used in combination with Cisplatin (ranked 5) as a first-line therapy for patients whose disease cannot be treated with surgery or radiation therapy [37], [38]. Specific KRAS mutations are responsible for lung cancer, and patients with these mutations are often resistant to targeted drugs such as those ranked 3 and 7 [39]. Moreover, drugs predicted to be active in NSCLC by the study in [35] were also present in our top-10 list (drugs ranked 4, 6, 8, 9, and 10), and have been validated by many studies [40]–[45].

Table 4. Structural patterns that describe drugs' activity in NSCLC obtained from our ML methods

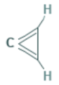
PubChem bit position	Definition of the descriptor	Chemical structure depiction	NSCLC activity
PubchemFP4	Presence of more than 1 Li molecule		highly active
PubchemFP165	Presence of more than 4 saturated or aromatic carbon-only ring of size 5		less active or inactive
PubchemFP50	Presence of more than 1 Mg molecule		highly active
PubchemFP51	Presence of more than 1 Al molecule		highly active
PubchemFP36	Presence of more than 8 S molecules		less active or inactive
PubchemFP104	Presence of more than 1 Sm molecule		highly active
PubchemFP119	Presence of at least one unsaturated non-aromatic carbon-only ring of size 3		highly active
PubchemFP171	Presence of more than 5 rings of size 5		less active or inactive
PubchemFP2	Presence of more than 16 H molecules		highly active
PubchemFP17	Presence of more than 8 N molecules		highly active
PubchemFP161	Presence of more than 3 unsaturated non-aromatic carbon-only rings of size 5		less active or inactive

Table 5. Top-10 ranked drugs in lung cancer

Rank	Drug name
1	Erlotinib
2	Paclitaxel
3	Atezolizumab
4	Osimertinib
5	Cisplatin
6	Fluoxetine
7	Sotorasib
8	Sulfasalazine
9	Azathioprine
10	Rotenone

5. CONCLUSION

In this paper, we propose a new approach to explore the important structures of active molecules in lung cancer. we set up two machine learning models to learn from chemical structures and predict novel drugs highly active in lung cancer, taking full advantage of a QSAR-based method. We conducted a comparative study to evaluate the performance of the two models based on several metrics. We used SHAP values to perform a feature engineering step and to list essential chemical structures that had a high contribution to the training phase of our models to make accurate predictions. Both models showed good results and were successfully able to rank the top 10 highly active molecules used as a therapy process for patients with lung cancer. The obtained results were compared to the medical research literature and supported by several studies. Our methodology demonstrated promising results that can enhance drug discovery pipelines not only for lung cancer case but can be generalized to other diseases.

REFERENCES

[1] T. Cheng, Q. Li, Z. Zhou, Y. Wang, and S. H. Bryant, "Structure-based virtual screening for drug discovery: a problem-centric review," *The AAPS Journal*, vol. 14, no. 1, pp. 133–141, Mar. 2012, doi: 10.1208/s12248-012-9322-0.

[2] H.-P. Shih, X. Zhang, and A. M. Aronov, "Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications," *Nature Reviews Drug Discovery*, vol. 17, no. 1, pp. 19–33, Jan. 2018, doi: 10.1038/nrd.2017.194.

[3] F. S. Hanoon and A. H. Hassin Alasadi, "A modified residual network for detection and classification of Alzheimer's disease," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 4400–4407, Aug. 2022, doi: 10.11591/ijece.v12i4.pp4400-4407.

[4] G. Saranya and A. Pravin, "A comprehensive study on disease risk predictions in machine learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 4, pp. 4217–4225, Aug. 2020, doi: 10.11591/ijece.v10i4.pp4217-4225.

[5] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, and A. Zhavoronkov, "Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data," *Molecular Pharmaceutics*, vol. 13, no. 7, pp. 2524–2530, Jul. 2016, doi: 10.1021/acs.molpharmaceut.6b00248.

[6] S. Kwon, H. Bae, J. Jo, and S. Yoon, "Comprehensive ensemble in QSAR prediction for drug discovery," *BMC Bioinformatics*, vol. 20, no. 1, Dec. 2019, doi: 10.1186/s12859-019-3135-4.

[7] S. Chackalamannil, D. Rotella, and S. E. Ward, *Comprehensive medicinal chemistry III*, 3rd Edition. Elsevier B.V. ScienceDirect, 2017.

[8] J. Vamathevan and E. Birney, "A review of recent advances in translational bioinformatics: Bridges from biology to medicine," *Yearbook of Medical Informatics*, vol. 26, no. 01, pp. 178–187, Sep. 2017, doi: 10.15265/IY-2017-017.

[9] S. Fotouhi, S. Asadi, and M. W. Kattan, "A comprehensive data level analysis for cancer diagnosis on imbalanced data," *Journal of Biomedical Informatics*, vol. 90, Feb. 2019, doi: 10.1016/j.jbi.2018.12.003.

[10] F. Rafii, B. D. R. Hassani, and M. A. Kbir, "Automatic clustering of microarray data using ART2 neural network," *Journal of Theoretical and Applied Information Technology*, vol. 90, no. 1, pp. 175–184, 2005.

[11] F. Rafii, M. A. Kbir, and B. D. R. Hassani, "MLP network for lung cancer presence prediction based on microarray data," in *2015 Third World Conference on Complex Systems (WCCS)*, Nov. 2015, pp. 1–6, doi: 10.1109/ICoCS.2015.7483276.

[12] H. Hanaf, B. Hassani, and M. Kbir, "Predicting gene-drug-disease interactions by integrating heterogeneous biological data through a network model," *International Journal of Advances in Soft Computing and its Applications*, vol. 14, no. 1, pp. 36–48, Apr. 2022, doi: 10.15849/IJASCA.220328.03.

[13] D. Altshuler, M. J. Daly, and E. S. Lander, "Genetic mapping in human disease," *Science*, vol. 322, no. 5903, pp. 881–888, Nov. 2008, doi: 10.1126/science.1156409.

[14] P. Ballesta, D. Bush, F. F. Silva, and F. Mora, "Genomic predictions using low-density SNP markers, pedigree and GWAS information: A case study with the non-model species eucalyptus cladocalyx," *Plants*, vol. 9, no. 1, Jan. 2020, doi: 10.3390/plants9010099.

[15] K. S. S. Lee *et al.*, "Drug-target residence time affects in vivo target occupancy through multiple pathways," *ACS Central Science*, vol. 5, no. 9, pp. 1614–1624, Sep. 2019, doi: 10.1021/acscentsci.9b00770.

[16] J. E. Larsen, T. Cascone, D. E. Gerber, J. V. Heymach, and J. D. Minna, "Targeted therapies for lung cancer," *The Cancer Journal*, vol. 17, no. 6, pp. 512–527, Nov. 2011, doi: 10.1097/PPO.0b013e31823e701a.

[17] X. Lu *et al.*, "The development of pharmacophore modeling: generation and recent applications in drug discovery," *Current Pharmaceutical Design*, vol. 24, no. 29, pp. 3424–3439, Dec. 2018, doi: 10.2174/1381612824666180810162944.

[18] S.-Y. Yang, "Pharmacophore modeling and applications in drug discovery: challenges and recent advances," *Drug Discovery Today*, vol. 15, no. 11–12, pp. 444–450, Jun. 2010, doi: 10.1016/j.drudis.2010.03.013.

[19] J. Almeida, J. Volpini, J. Poiani, and C. Silva, "Ligand-based drug design of novel MARK-3 inhibitors in cancer," *Current Bioactive Compounds*, vol. 10, no. 2, pp. 112–123, Oct. 2014, doi: 10.2174/1573407210021410011102743.

[20] M. Li, W. Zhang, B. Wang, Y. Gao, Z. Song, and Q. C. Zheng, "Ligand-based targeted therapy: a novel strategy for hepatocellular carcinoma," *International Journal of Nanomedicine*, vol. 11, pp. 5645–5669, Oct. 2016, doi: 10.2147/IJN.S115727.

[21] M. Akamatsu, "Current state and perspectives of 3D-QSAR," *Current Topics in Medicinal Chemistry*, vol. 2, no. 12, pp. 1381–1394, Dec. 2002, doi: 10.2174/1568026023392887.

[22] S. Vilar and G. Hripcsak, "The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug–drug interactions," *Briefings in Bioinformatics*, Jun. 2016, doi: 10.1093/bib/bbw048.

[23] R. Ferdousi, R. Safdari, and Y. Omidi, "Computational prediction of drug-drug interactions based on drugs functional similarities," *Journal of Biomedical Informatics*, vol. 70, pp. 54–64, Jun. 2017, doi: 10.1016/j.jbi.2017.04.021.

[24] H. Hanafi, B. D. R. Hassani, and M. A. Kbir, "Biological networks analysis, analytical approaches and use case on protein-protein network interactions," in *Proceedings of the 4th International Conference on Smart City Applications*, Oct. 2019, pp. 1–5, doi: 10.1145/3368756.3368996.

[25] C.-H. Huang, P. M.-H. Chang, C.-W. Hsu, C.-Y. F. Huang, and K.-L. Ng, "Drug repositioning for non-small cell lung cancer by using machine learning algorithms and topological graph theory," *BMC Bioinformatics*, vol. 17, no. S1, Dec. 2016, doi: 10.1186/s12859-015-0845-0.

[26] C.-W. Weng *et al.*, "Pharmacophore-based virtual screening for the identification of the novel Src inhibitor SJG-136 against lung cancer cell growth and motility," *American Journal of Cancer Research*, vol. 10, no. 6, pp. 1668–1690, 2020.

[27] A. Gaulton *et al.*, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, no. D1, pp. D1100–D1107, Jan. 2012, doi: 10.1093/nar/gkr777.

[28] A. Gaulton *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Research*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, doi: 10.1093/nar/gkw1074.

[29] C. W. Yap, "PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints," *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1466–1474, May 2011, doi: 10.1002/jcc.21707.

[30] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/ci00057a005.

[31] N. Ketkar, "Introduction to keras," in *Deep Learning with Python*, Berkeley, CA: Apress, 2017, pp. 97–111, doi: 10.1007/978-1-4842-2766-4_7.

[32] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[33] P. D. Chary and R. P. Singh, "Review on advanced machine learning model: scikit-learn," *International Journal of Scientific Research and Engineering Development*, vol. 3, no. 4, pp. 526–529, 2020. Accessed: Sep. 20, 2022. [Online]. Available: https://papers.ssrn.com/abstract=3694350.

[34] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 4768–4777.

[35] R. X. Huang *et al.*, "Lung adenocarcinoma-related target gene prediction and drug repositioning," *Frontiers in Pharmacology*, vol. 13, Aug. 2022, doi: 10.3389/fphar.2022.936758.

[36] C. M. Rocha-Lima and L. E. Raez, "Erlotinib (tarceva) for the treatment of non-small-cell lung cancer and pancreatic cancer," *P & T: A Peer-reviewed Journal for Formulary Management*, vol. 34, no. 10, pp. 559–564, 2009.

[37] S. Ramalingam and C. P. Belani, "Paclitaxel for non-small cell lung cancer," *Expert Opinion on Pharmacotherapy*, vol. 5, no. 8, pp. 1771–1780, Aug. 2004, doi: 10.1517/14656566.5.8.1771.

[38] R. Pirker, G. Krajnik, S. Zöchbauer, R. Malayeri, M. Kneussl, and H. Huber, "Paclitaxel/cisplatin in advanced non-small-cell lung cancer (NSCLC)," *Annals of Oncology*, vol. 6, no. 8, pp. 833–835, Oct. 1995, doi: 10.1093/oxfordjournals.annonc.a059324.

[39] L. Huang, Z. Guo, F. Wang, and L. Fu, "KRAS mutation: from undruggable to druggable in cancer," *Signal Transduction and Targeted Therapy*, vol. 6, no. 1, Nov. 2021, doi: 10.1038/s41392-021-00780-4.

[40] X. Tang *et al.*, "Machine learning-based CT radiomics analysis for prognostic prediction in metastatic non-small cell lung cancer patients with EGFR-T790M mutation receiving third-generation EGFR-TKI Osimertinib treatment," *Frontiers in Oncology*, vol. 11, Sep. 2021, doi: 10.3389/fonc.2021.719919.

[41] K. Hu *et al.*, "Suppression of the SLC7A11/glutathione axis causes synthetic lethality in KRAS-mutant lung adenocarcinoma," *Journal of Clinical Investigation*, vol. 130, no. 4, pp. 1752–1766, Mar. 2020, doi: 10.1172/JCI124049.

[42] Y.-L. Shi, S. Feng, W. Chen, Z.-C. Hua, J.-J. Bian, and W. Yin, "Mitochondrial inhibitor sensitizes non-small-cell lung carcinoma cells to TRAIL-induced apoptosis by reactive oxygen species and Bcl-XL/p53-mediated amplification mechanisms," *Cell Death & Disease*, vol. 5, no. 12, pp. e1579–e1579, Dec. 2014, doi: 10.1038/cddis.2014.547.

[43] Z. Yang *et al.*, "Antitumor effect of fluoxetine on chronic stress-promoted lung cancer growth via suppressing kynurenine pathway and enhancing cellular immunity," *Frontiers in Pharmacology*, vol. 12, Aug. 2021, doi: 10.3389/fphar.2021.685898.

[44] D. A. Gomes, A. M. Joubert, and M. H. Visagie, "In vitro effects of papaverine on cell proliferation, reactive oxygen species, and cell cycle progression in cancer cells," *Molecules*, vol. 26, no. 21, Oct. 2021, doi: 10.3390/molecules26216388.

[45] K. W. Baik, S. H. Kim, and H. Y. Shin, "Paraneoplastic neuromyelitis optica associated with lung adenocarcinoma in a young woman," *Journal of Clinical Neurology*, vol. 14, no. 2, pp. 246–247, 2018, doi: 10.3988/jcn.2018.14.2.246.

BIOGRAPHIES OF AUTHORS

Hamza Hanafi received his degree in Computer Science engineering in 2017 from the Faculty of Sciences and Technologies at the University Abdelmalek Essaâdi in Tangier, Morocco. He is currently a Ph.D. candidate and member of the Intelligent Automation Team at the STI Doctoral Center, also located in the Faculty of Sciences and Technologies at the University Abdelmalek Essaâdi. His research interests include computational biology, bioinformatics, and machine learning. Hamza has published several research articles in international journals and conferences on computer science. He can be contacted at email: hamzahanafi1@gmail.com or hamza.hanafi@etu.uae.ac.ma.

Badr Dine Rossi Hassani is a full professor of biology at the Faculty of Sciences and Technologies (FST) of Tangier, Morocco, and Ph.D. managing director at LABIPHABE laboratory, his research areas interest many disciplines: Cancer Research, Biotechnology, Bioinformatics. He is a member of scientific committees of many international conferences and journals. He can be contacted at email: badrrossi@gmail.com.

M'hamed Aït Kbir is a full professor at the computer science department of the Faculty of Sciences and Technologies (FST) of Tangier, since 2001, University Abdelmalek Essaâdi, Morocco. As a member of LIST laboratory, since 2007, his research works focus on three main areas: Computer vision (multimedia flow optimization, multimedia document content watermarking, object recognition, 3D contents indexing and retrieval, 3D reconstruction)-artificial intelligence (machine learning, deep learning, planning and search strategies)-bioinformatics (micro-array data decision making, biological data integration, biological networks analysis). He is a member of scientific committees of many international conferences and journals. As an expert, he participates in the evaluation of public and private education programs for the ANEAQ and the ministry of higher education and scientific research. He can be contacted by email: maitkbir@uae.ac.ma.