

Non-small cell lung cancer active compounds discovery holding on protein expression using machine learning models

Hamza Hanafi¹, M'hamed Aït Kbir¹, Badr Dine Rossi Hassani²

¹Intelligent Automation and BioMedGenomics Laboratory, STSM Doctoral Center, Abdelmalek Essaadi University, Tangier, Morocco

²LABIPHABE Laboratory, STI Doctoral Center, Abdelmalek Essaadi University, Tangier, Morocco

Article Info

Article history:

Received May 23, 2024

Revised Feb 25, 2025

Accepted Mar 15, 2025

Keywords:

Drug discovery

Lung cancer

Machine learning models

Precision medicine

Protein expressions

ABSTRACT

Computational methods have transformed the field of drug discovery, which significantly helped in the development of new treatments. Nowadays, researchers are exploring a wide range of opportunities to identify new compounds using machine learning. We conducted a comparative study between multiple models capable of predicting compounds to target non-small cell lung cancer, we focused on integrating protein expressions to identify potential compounds that exhibit a high efficacy in targeting lung cancer cells. A dataset was constructed based on the trials available in the ChEMBL database. Then, molecular descriptors were calculated to extract structure-activity relationships from the selected compounds and feed into several machine learning models to learn from. We compared the performance of various algorithms. The multilayer perceptron model exhibited the highest F1 score, achieving an outstanding value of 0,861. Moreover, we present a list of 10 drugs predicted as active in lung cancer, all of which are supported by relevant scientific evidence in the medical literature. Our study showcases the potential of combining protein expression analysis and machine learning techniques to identify novel drugs. Our analytical approach contributes to the drug discovery pipeline, and opens new opportunities to explore and identify new targeted therapies.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Hamza Hanafi

Intelligent Automation and BioMedGenomics Laboratory, STSM Doctoral Center

Abdelmalek Essaadi University

Tangier, Morocco

Email: hamza.hanafi@etu.uae.ac.ma

1. INTRODUCTION

Drug discovery plays a fundamental role in the healthcare sector, as developing new compounds demands a multidisciplinary approach to provide novel therapeutic interventions. Despite this, the process is often complex, time-consuming, and requires an enormous effort to validate new treatments. Moreover, traditional methods of drug discovery are not only resource-intensive but also limited in their scope [1].

Recent advancements in computational biology have completely transformed drug discovery pipelines. The combination of biology with computational methods offers new insights to accelerate the identification and evaluation of novel compounds. Therefore, computational techniques have emerged as powerful tools in the field of pharmacological medicine [2], and revealed great success compared to traditional methods. Besides, these techniques have found widespread application in various healthcare domains, including disease classification [3] and surgical enhancements [4].

Nowadays, a large amount of biological data is stored in public databases and enables researchers to explore a wide range of methodologies. Furthermore, the integration and analysis of this biological data ease

the study of new hypotheses [5], for example, predictive modeling using machine learning (ML) techniques is one of the most explored methodologies and has gained prominence. ML models can effectively classify drugs into relevant therapeutic categories, accurately detect and classify tumor stages [6], and design new drugs based on chemical properties [7]. Consequently, ML-based methods are capable in detecting patterns and identifying correlations within large and complex datasets with numerous variables.

Furthermore, bioinformatic methods have been crucial in the drug discovery pipelines, allowing researchers to study molecules from a system-level perspective. By integrating knowledge from various domains such as genomics, proteomics, transcriptomics, population genetics, and molecular phylogenetics, bioinformatic analysis eases drug target identification, drug candidate screening, prediction of drug resistance, and minimization of side effects. Thus, ML algorithms are employed alongside bioinformatics to predict interactions among biological entities [8] and design customized drugs for specific treatments, ultimately advancing precision medicine.

However, researchers have to face several challenges to build ML models in drug discovery. Biological data often varies in quality and always needs preprocessing before it can be used for learning purposes [9]. Additionally, cancer classification problems typically involve imbalanced datasets, characterized by both excessive noise and a deficiency of labeled data, which significantly affects the learning process. Evaluating the efficacy of ML models in such scenarios becomes complex, particularly when confronted with limited or biased data [10].

Our contribution aims to develop a ML-based classifier capable of predicting active compounds that can target non-small cell lung cancer (NSCLC). First, we curated a dataset by extracting bioactivity data from ChEMBL [11] database based on proteins expressed in NSCLC. Second, molecular descriptors of the selected compounds were calculated and used as input feature for several models. Then, numerous ML models were fed with this data and trained to learn from the structure and chemical characteristics of the molecules. Finally, we performed a comparative analysis to identify the optimal model.

The rest of the paper is organized as follows: section 2 provides an overview of related works in the field. Section 3 presents our approach, including data collection and the methodology employed. Section 4 discusses the results obtained from our analysis, followed by the conclusion in section 5.

2. RELATED WORK

Nowadays, many studies have been conducted to explore and understand the biological aspects of cancer cells using ML models. In particular, these studies aim to better explain the mechanisms of different signaling pathways that transmit signals within cells and affect genes regulation. Proteins such as Ras play an important role in regulating various biomolecular interactions in the cell's lifecycle [12]. The Ras pathways transmit signals to activate genes that promote cell growth and division. Mutations in genes associated with these pathways can lead to different types of cancers [13], [14]. Therefore, there is a growing interest in identifying new anti-Ras therapeutic strategies.

In a study conducted by Way *et al.* [15], three types of biological data were explored: gene expressions, mutation counts, and mutation copies found in various types of cancers using ML methods to predict the activation of the Ras pathways. The authors of this study were able to design a model capable of predicting RNA sequences that activate the Ras pathways. Similarly, Knijnenburg *et al.* [16] employed genomic and molecular data to predict the activation of p53 pathways. The gene TP53 contains instructions for regulating a protein called p53, which functions as a tumor suppressor and interacts with the apoptosis mechanism [17]. Consequently, mutations in the gene TP53 can lead to metastatic cancer [18].

Some metastatic cancers are associated with the loss of phenotypic traits expressed by stem cells [19]. In this context, to elucidate the relationship between tumor differentiation phenotype and tumor propagation or genetic alterations, Malta *et al.* [20] introduced an ML model aimed at predicting cancer development within specific cellular tissues. They relied on data from stem cells and their progenitor cells to construct a classifier for genetic expression traits. Subsequently, they applied this classifier to a cell sample to predict the expressed traits. They were able to identify cancer cells within the sample, but they did not provide detailed information about the learning methodology used to build the classifier.

Mutations in the epidermal growth factor receptor (EGFR) have been known to cause uncontrolled cell proliferation [21]. Numerous studies aim to identify small inhibitory molecules that target the EGFR gene. Qureshi *et al.* [22] proposes a personalized model for predicting drug response in lung cancer patients. Specifically, this model was tested to predict the response to US Food and Drug Administration (FDA)-approved small molecules, such as Erlotinib and Gefitinib. To construct their model, the authors assembled various types of data: EGFR mutations found in lung cancer patients, clinical data including patient survival and clinical response to drugs, demographic data such as age, sex, and smoking history, and the 3D structure of EGFR gene mutations found in patients.

A decision tree-based classifier was trained using this data to predict the level of drug response among four categories: no response, partial response, moderate response, and strong response. The authors found that demographic data had a weak impact on the learning outcome of the model. Only EGFR mutations and structures showed a good predictive response level of drug response. The authors did not further use this model to test the response level of molecules that were not used during the learning phase. Yang *et al.* [23] aiming to determine the data that can establish an ML model to predict EGFR mutations in lung cancer, the authors compared the performance of several learning algorithms random forest (RF), light gradient boosting machine (LightGBM), support vector machine (SVM), multilayer perceptron (MLP), and extreme gradient boosting (XGB) using multiple clinical and demographic data. They found that tobacco consumption, sex, cholesterol, age, and the albumin/globulin ratio were among the top five variables related to EGFR mutation, which differed slightly from the results obtained in the study [22], where the impact of demographic data was weak.

Widyananda *et al.* [24] investigate the potential of Quercetin, a natural compound found in fruits and vegetables, to combat glioblastoma multiforme. By examining databases like national center for biotechnology information (NCBI), super-enhancer archive (SEA), comparative toxicogenomics database (CTD), and search tool for the retrieval of interacting genes/proteins (STRING), the study identifies four key proteins serine/threonine kinase 1 (AKT1), matrix metalloproteinase 9 (MMP9), ATP binding cassette subfamily B member 1 (ABCB1), and vascular endothelial growth factor A (VEGFA), that Quercetin directly affects. Using STITCH, SEA, and STRING, the study constructs protein-protein interaction networks, highlighting connections between these proteins. Functional annotation analysis through the DAVID web server clarifies the biological processes influenced by these proteins. Molecular docking simulations with AutoDock Vina [25] provide insights into how Quercetin interacts with these proteins, extending our understanding of its potential as a glioblastoma multiforme treatment. The study not only uncovers Quercetin's impact on crucial glioblastoma multiforme related proteins but also emphasizes its potential as a targeted therapeutic option against glioblastoma multiforme.

3. METHODOLOGY

Quantitative structure-activity relationship (QSAR) modeling leverages the relationships between the chemical structure and the biological activity of molecules [26]. QSAR models employ molecular descriptors, which capture the physical and chemical properties distinguishing one molecule from another [27]. These models provide valuable insights into the chemical properties that are crucial for the inhibition of specific biological processes. Thus, aiding biologists and chemists in the design of robust molecules with optimized properties. Utilizing ML-based QSAR analysis and molecular docking, Iresha *et al.* [28] explores medicinal plant compounds as inhibitors for HIV-1 reverse transcriptase, addressing resistance issues. Similarly, our study aims to use ML models to predict active compounds based on the genes expressed in NSCLC through QSAR analysis.

To construct our dataset, First, we selected a set of genes that have been extensively associated with NSCLC in various studies [6], [7]. After gathering the target proteins related to these genes from the ChEMBL database, we selected their bioactivities and computed their molecular descriptors to analyze the chemical structure and identify patterns in active compounds. Afterwards, we trained several models using the constructed descriptors and evaluated their performance based on the confusion matrix and the achieved F1 score. These evaluation metrics provide a comprehensive assessment of the models' predictive abilities.

Figure 1 illustrates our proposed methodology, and the experimental procedure established. The term "targets" in the ChEMBL database refers to proteins or organisms that compounds act upon. Biologically, these compounds engage in interactions with the targeted proteins, resulting in a modulatory activity. Such activity may encompass the activation or inhibition of the targeted protein. The overall approach is followed to predict compounds' activity to target NSCLC.

In step 1 the expressed genes are identified from the medical literature [29], this includes 8 expressed genes: b-raf proto-oncogene, serine/threonine kinase (BRAF), EGFR, kirsten rat sarcoma virus–proto-oncogene, GTPase (KRAS), phosphatase and tensin homolog (PTEN), receptor tyrosine kinase (ROS1), v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2, also known as HER2 and neu (ERBB2), MET proto-oncogene, receptor tyrosine kinase (MET), and anaplastic lymphoma kinase (ALK). In step 2, bioactivity data of the target protein is extracted from ChEMBL database. In step 3, molecular descriptors of the bioactivity data are calculated. In step 4, several ML models are trained on these molecular descriptors. Finally, in step 5, the models are evaluated to assess their predictive accuracy for determining compound activity to target NSCLC.

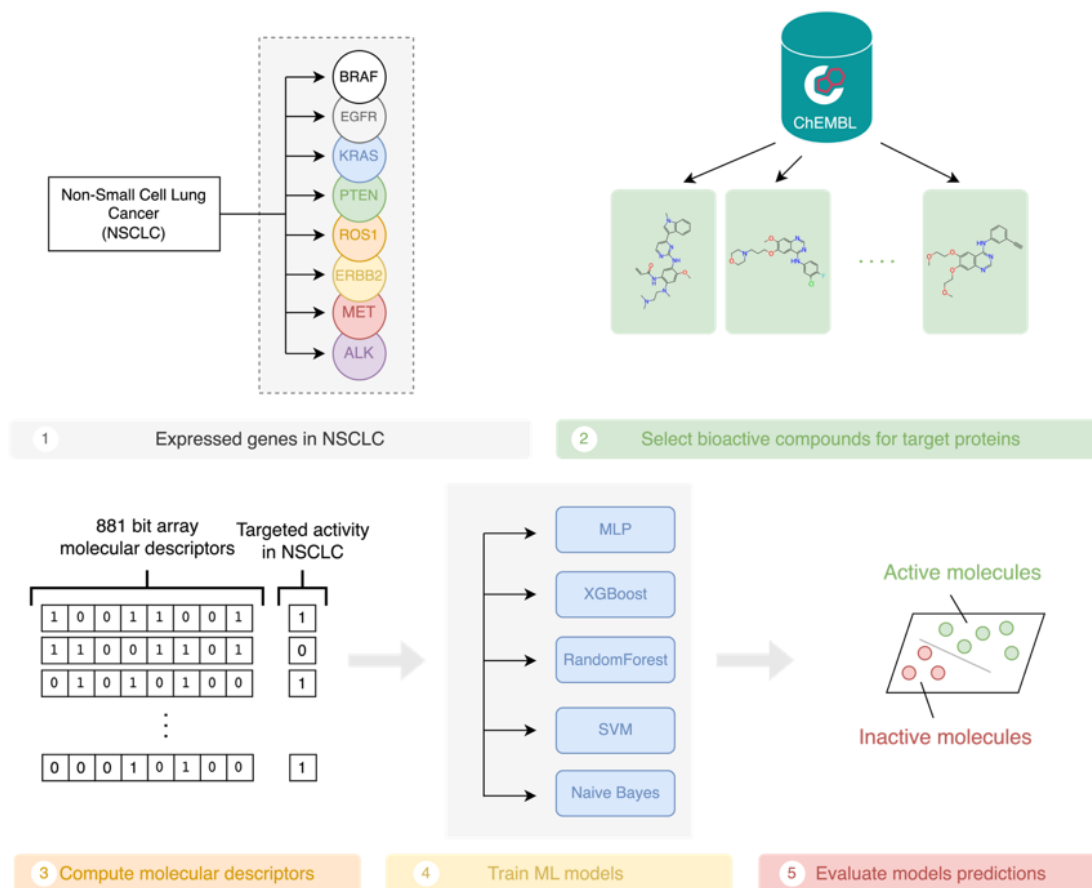


Figure 1. Proposed methodology for predicting compound activity to target NSCLC using bioactivity data, molecular descriptors, and ML models

3.1. Molecular descriptors

Molecular descriptors are numerical representations that capture various physicochemical and topological properties of molecules. These descriptors provide valuable quantitative information about the characteristics and behavior of chemical compounds. By analyzing and comparing molecular descriptors, researchers can gain insights into the structure-activity relationships of molecules and make predictions about their properties, reactivity, and potential biological activities.

One commonly used tool for computing molecular descriptors is PaDEL-descriptor [30]. It is a software program that calculates a comprehensive set of molecular descriptors based on simplified molecular input line entry system (SMILES) notations [31]. SMILES is a compact string representation of a molecule's structure, which encodes key structural features including atom types, bond connections, and their spatial arrangement within the molecule. The PaDEL-descriptor utilizes algorithms and mathematical formulas to generate a wide range of descriptors, including constitutional, topological, and physicochemical descriptors.

Constitutional descriptors capture basic molecular features, such as the number of atoms, bonds, and functional groups. Topological descriptors assess molecular connectivity and shape, providing information about the arrangement of atoms and the presence of specific structural motifs. Physicochemical descriptors quantify properties such as molecular weight, solubility, lipophilicity, hydrogen bonding potential, and electronic properties. In our studying, we used the descriptors defined by the PubChem database [32], which primarily focus on the structural and physicochemical properties of compounds. These descriptors are typically encoded in a byte array. Table 1 provides a description of the PubChem descriptors bytes.

3.2. Learning tasks

To construct our dataset, 84,078 molecules were selected from ChEMBL database and classified them into active and inactive sets based on their inhibition concentration value at 50% (IC₅₀). Initially, we used a

threshold of 100 nM, and after iterative training of the ML models with varying thresholds, we observed that lowering the IC₅₀ threshold improved the models' predictive performance. Further experimentation revealed that setting the threshold at 77 nM yielded the highest F1 score, indicating optimal model performance that allows the model to effectively explore the structure of compounds to determine its activity. Thus, choosing this threshold was crucial to enhance the predictive capability of the ML algorithms used in this study. Consequently, molecules with an IC₅₀ value lower than or equal to 77 nM were considered active, denoted by an assigned activity level of 1, while those greater than 77 nM were considered inactive, represented by an activity level of 0.

Table 1. Summary description of PubChem descriptors

PubChem bit position range	Description
From 0 to 114	These binary units examine the presence or abundance of specific chemical atoms.
From 115 to 262	These binary units assess the presence of cyclic structures.
From 263 to 326	These binary units examine the presence of connected pairs of atoms, disregarding their quantity and arrangement.
From 327 to 448	These binary units assess the presence of atom nearest neighbor patterns, considering the relevance of aromaticity and significant bonding.
From 445 to 459	These binary units examine complex atom neighborhood patterns, irrespective of their quantity, with specific consideration given to bond orders.
From 460 to 712	These binary units evaluate the presence of straightforward SMILES arbitrary target specification (SMARTS) patterns, without considering their quantity, but with specific attention given to bond orders and the compatibility of bond aromaticity with both single and double bonds.
From 713 to 880	These binary units examine the presence of complex SMARTS patterns, irrespective of their quantity, with particular emphasis on specific bond orders and bond aromaticity.

To train our models, 90% of the data was allocated to the training set, while the remaining 10% was used for the test set. Molecular descriptors are going to serve as input features, while the target feature is the activity level in NSCLC. Given the complexity introduced by this extensive array of input features, we implemented a preprocessing step to refine the dataset, to make the ML models more precise and to depict the patterns of the most important molecular descriptors. In this regard, before feeding the data into the ML models, we performed an initial step to reduce the number of input features. Initially, there were 881 features; by applying a variance threshold of 0.16, we removed molecular descriptors with low variance, which showed 84% similarity in their values. This resulted in a final set of 160 features with higher variance, enabling the model to detect meaningful patterns within the dataset.

We trained several ML models on the computed molecular descriptors, including MLP, XGB, RF, SVM, and naive Bayes (NB). Firstly, the MLP neural network model is used with an input layer of 100 units and uses the rectified linear unit (ReLU) as an activation function. This is followed by some hidden layers with 50, 20, and 5 units, respectively, also using the ReLU activation function. And the output layer has a single unit with a sigmoid activation function, allowing for binary classification. During training, we utilized the Adam optimizer with a learning rate of 0.001 and employed the binary cross-entropy as a loss function. The model was trained for 100 epochs with a batch size of 100 samples. Furthermore, the training data was split into training and validation subsets, with a 5% validation split.

We also trained two tree-based classifiers, including XGB and RF models. These models, constructed using the sklearn implementation, are based on ensemble learning techniques that combines multiple decision trees to make predictions. To construct the models, we used various parameters to optimize their performance. We used a max depth of 2 levels, for each constructed individual decision tree. Additionally, we utilized a learning rate of 0.01, to control the step size at each boosting and bagging iteration. The number of estimators was set to 1,000, indicating the number of decision trees to be created in the ensemble. This value was chosen from a list that included 50, 200, 400, 600, 800, and 1,000 estimators. The receiver operating characteristic (ROC) curves were subsequently drawn for each model trained with a distinct number of estimators to assess their impact on the model's performance. Notably, as the number of estimators increased from 50 to 1,000, we observed a progressive improvement of the area under the curve (AUC). The achieved AUC values stood at 0.621 for XGB and 0.596 for RF as shown in Figures 2 and 3 respectively. We also adjusted the class weights to address class imbalance and enhance the model's performance, particularly on the minority class by assigning a higher weight to samples from the minority class.

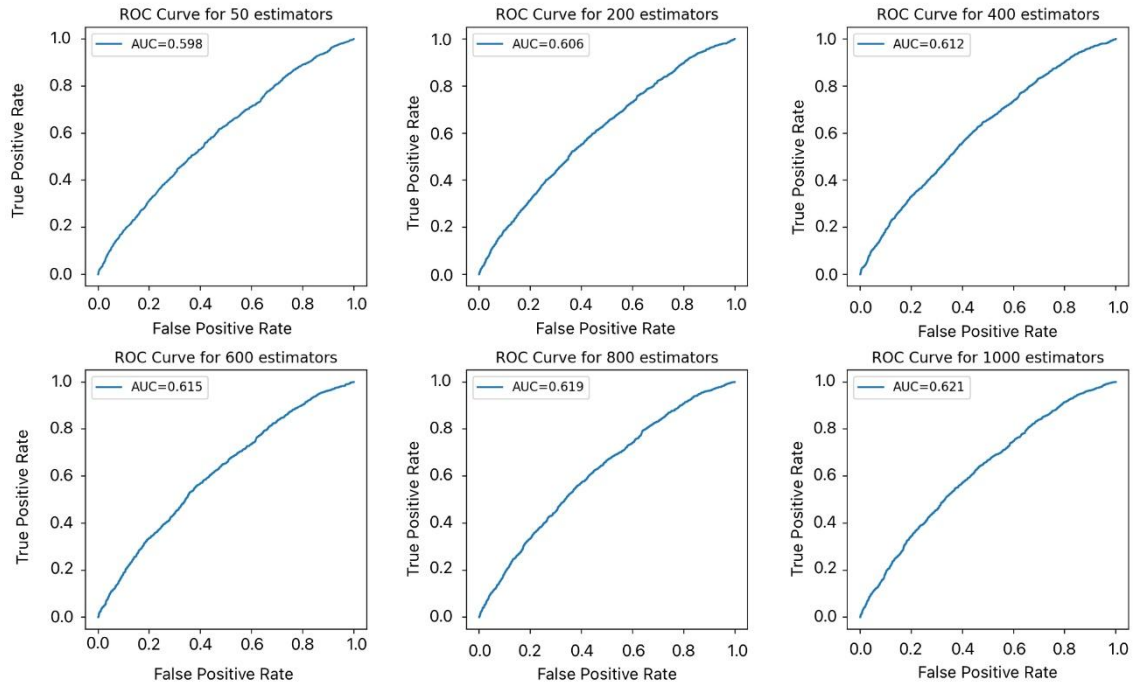


Figure 2. ROC curves for XGB with varied numbers of estimators

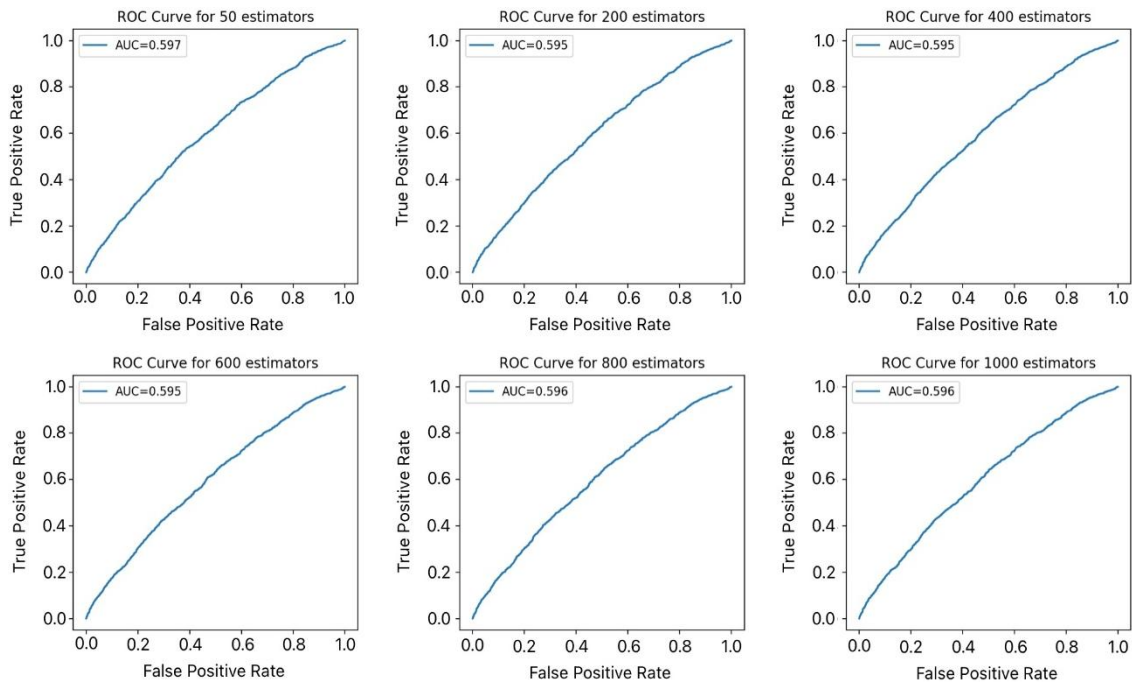


Figure 3. ROC curves for RF with varied numbers of estimators

Furthermore, we built an SVM model using different kernel functions. This includes linear, radial basis function (RBF), polynomial, and sigmoid functions shown in Figure 4. It appears that for linear, RBF, and polynomial kernels, the training score is relatively high when using few samples for training and decreases when increasing the number of samples. In contrast, the cross-validation score starts at a moderate level and shows a slight increase when adding samples. Whereas the plot for Sigmoid kernel, the training score remains low regardless of the size of the training set. On the other hand, the cross-validation score

decreases with the size of the training dataset. Indeed, it decreases to a point where it reaches a plateau. The polynomial and RBF kernels enables us to classify the data with complex relationships. The model was trained to find the best boundary that separates the active and inactive molecules. We set the regularization parameter to 10 to strike a good balance between training accuracy and classification precision, and we used the 'scale' option for gamma to ensure a smooth decision boundary. These choices allow our SVM model to perform effectively, overcoming the challenge of class imbalance inherent in the data. Lastly, we utilized the sklearn library to implement a NB classifier, which is a probabilistic classifier that assumes feature independence, making it efficient and suitable for large datasets. The default implementation in sklearn employs the Gaussian NB algorithm, assuming a Gaussian distribution for the features.

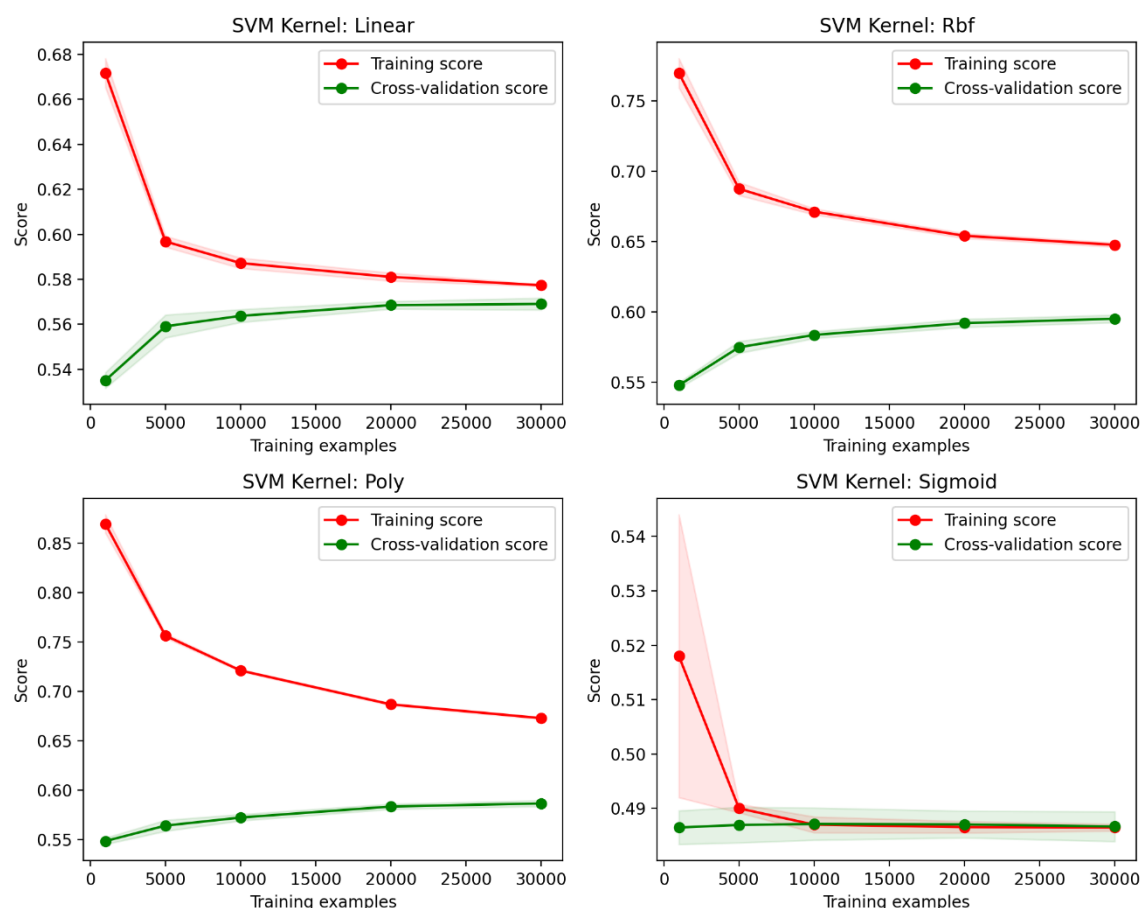


Figure 4. Learning curves for SVM with different kernels (RBF, poly, linear)

4. RESULTS AND DISCUSSION

This section presents the key findings from our study, which focuses on comparing different algorithms to find the best model capable of predicting the activity of compounds targeting a specific protein. Although numerous models exist in the literature, there is a notable gap in identifying the optimal one. Our study addresses this gap by evaluating various models using performance metrics. Table 2 provides a comprehensive summary of the overall accuracy, precision, recall, and F1 score. In contrast, Figure 5 displays a heatmap illustrating the accuracy, precision, and recall achieved by each model across different classes.

Table 2. Overall performance metrics of trained ML models

Model	Accuracy	Precision	Recall	F1 score
MLP	0.865	0.878	0.845	0.861
XGB	0.604	0.601	0.605	0.603
RF	0.564	0.551	0.658	0.600
SVM	0.593	0.595	0.598	0.597
NB	0.558	0.546	0.649	0.593



Figure 5. Heatmap showcasing the accuracy, precision, and recall scores attained by each model

The MLP model achieved an accuracy of 0.865, indicating that it correctly predicted the activity of compounds for inhibiting NSCLC in a large majority of cases. It had a precision of 0.878, meaning that when it predicted a compound as effective against NSCLC it demonstrated the model's ability to exclude irrelevant cases. The F1 score of 0.861, which is the harmonic mean of precision and recall, indicates the good overall balanced performance of the model. These results suggest that the MLP model performed the best among the tested models in predicting the activity of compounds for inhibiting NSCLC.

The XGB model achieved an accuracy of 0.604, indicating moderate performance in predicting the activity of compounds for inhibiting NSCLC. It had a precision of 0.601, recall of 0.605 and an F1 score of 0.603, suggesting a relatively balanced performance. While the accuracy of the XGB model is lower compared to the MLP model, it still provides a reasonable level of predictive ability in identifying potential compounds for NSCLC inhibition.

The RF model achieved an accuracy of 0.564, which is lower than the MLP and XGB models. It had a precision of 0.551, indicating a higher rate of false positives compared to the other models. However, the recall value of 0.658 suggests that the RF model successfully identified a higher proportion of true positives (compounds with NSCLC inhibitory activity) compared to other models. The F1 score of 0.600 reflects a moderately balanced performance between precision and recall. Overall, while the RF model shows potential in capturing true positive cases, it suffers from a higher rate of false positives, impacting its overall accuracy in predicting compounds for NSCLC inhibition.

The SVM and NB models exhibit weaker performances compared to more advanced models, such as MLP. The SVM model achieved an accuracy of 0.593 with a balanced precision of 0.595, recall of 0.598, and an F1 score of 0.597. While the SVM model demonstrates potential in capturing compounds with and without NSCLC inhibitory activity, its accuracy falls below that of the MLP model, suggesting limitations in accurately predicting compound activity. Similarly, the NB model demonstrates a balance between precision of 0.546 and recall of 0.649, resulting in a moderately balanced F1 score of 0.593. The NB model, while providing some predictive ability, lags behind other models in terms of accuracy and precision. This analysis highlights the challenges faced by both SVM and NB models in effectively capturing the complexities of the data for accurate predictions in NSCLC inhibition.

Among these models, the MLP model performed the best with the highest accuracy of 0.865, indicating its ability to make accurate predictions. It also achieved the highest precision and F1 score values. On the other hand, the NB model performed the least with an accuracy of 0.558, indicating a lower level of prediction accuracy compared to the other models. It had the lowest precision and F1 score values, suggesting its limitations in accurately predicting the activity for NSCLC. Using the MLP model, we ranked top-10 highly active molecules in NSCLC. Table 3 shows a list of these drugs. The ranking method is based on the probabilities returned by the MLP Model, where these probabilities represent the percentage of belonging to the positive class, demonstrating the likelihood of a compound's activity against NSCLC.

The list of drugs predicted by our MLP model to target NSCLC aligns well with the drugs mentioned in the medical literature. Several of the drugs in the top-10 list, such as Osimertinib, Brigatinib, Alectinib, Erlotinib, Ceritinib, Afatinib, Trastuzumab, Adagrasib and Gefitinib that are recognized as important therapeutic agents for NSCLC [33]. The literature highlights the effectiveness of these drugs in various settings, including advanced NSCLC with specific genetic mutations (such as EGFR mutations, ALK-positive or ROS-1-positive NSCLC) and metastatic NSCLC. The literature also provides supporting

evidence for the efficacy of these drugs, with information on overall survival, and improved survival time [33]. Moreover, the fact that some of the drugs in the top-10 list are approved such as Osimertinib [34], Brigatinib [35], Alectinib [36], Erlotinib [37], Ceritinib [38], Afatinib [39], and Gefitinib [40], underscores their established efficacy in NSCLC treatment. While other drugs such as Sotorasib [41], Trastuzumab [42], and Adagrasib [43] are currently undergoing clinical trials further confirms their relevance in NSCLC treatment.

Table 3. Top-10 ranked drugs in lung cancer

Rank	Drug name
1	Osimertinib
2	Brigatinib
3	Alectinib
4	Erlotinib
5	Ceritinib
6	Afatinib
7	Sotorasib
8	Trastuzumab
9	Adagrasib
10	Gefitinib

5. CONCLUSION

Our study showed the potential of integrating protein expression analysis and ML techniques for active compounds discovery in lung cancer treatment. By leveraging gene expression data and targeted protein analysis, we successfully identified bioactive compounds that specifically target proteins associated with NSCLC. Through using various ML models, including MLP, XGB, RF, SVM, and NB, we compared their performances in predicting the activity of compounds. Among these models, the MLP model exhibited the highest F1 score, achieving an impressive value of 0.861, denoting its ability to accurately predict active compounds for NSCLC treatment. Furthermore, our study provides a list of 10 drugs predicted as active in NSCLC, all of which are supported by relevant scientific evidence. These findings contribute to the drug discovery pipeline for lung cancer, offering valuable insights into the development of targeted therapies. Accordingly, the integration of computational methods with bioinformatic tools provides a powerful approach to accelerate the identification and evaluation of novel compounds, ultimately advancing precision medicine in the treatment of lung cancer. Future research will focus on validating the predicted compounds in preclinical and clinical studies to further confirm their efficacy.

ACKNOWLEDGEMENTS

The authors express their gratitude to the reviewers for their constructive feedback during the development of this work. This research was conducted independently, without external funding or financial support.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hamza Hanafi	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓			
M'hamed Aït Kbir	✓	✓		✓			✓		✓	✓		✓	✓	
Badr Dine Rossi	✓	✓		✓			✓			✓		✓	✓	
Hassani														

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nvestigation

R : **R**esources

D : **D**ata Curation

O : **O**riting - **O**riginal Draft

E : **E**riting - **R**eview & **E**editing

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study were obtained from the ChEMBL database, which is publicly available at <https://www.ebi.ac.uk/chembl>.




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," *AAPS Journal*, vol. 14, no. 1, pp. 133–141, 2012, doi: 10.1208/s12248-012-9322-0.
- [2] M. M. Rahman *et al.*, "Emerging promise of computational techniques in anti-cancer research: at a Glance," *Bioengineering*, vol. 9, no. 8, 2022, doi: 10.3390/bioengineering9080335.
- [3] G. Hussain and Y. Shiren, "Identifying Alzheimer disease dementia levels using machine learning methods," *Medical Research Archives*, vol. 11, no. 7.1, 2023, doi: 10.18103/mra.v11i7.1.4039.
- [4] M. Sugimoto and T. Sueyoshi, "Development of holoeyes holographic image-guided surgery and telemedicine system: clinical benefits of extended reality (virtual reality, augmented reality, mixed reality), the metaverse, and artificial intelligence in surgery with a systematic review," *Medical Research Archives*, vol. 11, no. 7.1, 2023, doi: 10.18103/mra.v11i7.1.4045.
- [5] A. H. Urbanski, J. D. Araujo, R. Creighton, and H. I. Nakaya, "Integrative biology approaches applied to human diseases," *Computational Biology*, pp. 19–36, 2019, doi: 10.15586/computationalbiology.2019.ch2.
- [6] A. K. AAIAbdulsalam, J. H. Garvin, A. Redd, M. E. Carter, C. Sweeny, and S. M. Meystre, "Automated extraction and classification of cancer stage mentions from unstructured text fields in a central cancer registry," *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, vol. 2017, pp. 16–25, 2018.
- [7] H. Hanafi, B. D. R. Hassani, and M. A. Kbir, "Predicting active compounds for lung cancer based on quantitative structure-activity relationships," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 5, pp. 5755–5763, 2023, doi: 10.11591/ijece.v13i5.pp5755-5763.
- [8] H. Hanafi, B. D. R. Hassani, and M. A. 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. 35–48, 2022, doi: 10.15849/IJASCA.220328.03.
- [9] A. Sivakumar and R. Gunasundari, "A survey on data preprocessing techniques for bioinformatics and web usage mining," *International Journal of Pure and Applied Mathematics*, vol. 117, no. 20, pp. 785–794, 2017.
- [10] Y. Zhang and J. Hong, "Challenges of deep learning in cancers," *Technology in Cancer Research and Treatment*, vol. 22, 2023, doi: 10.1177/15330338231173495.
- [11] A. Gaulton *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Research*, vol. 45, no. 1, pp. 945–954, 2017, doi: 10.1093/nar/gkw1074.
- [12] D. K. Simanshu, D. V. Nissley, and F. McCormick, "RAS proteins and their regulators in human disease," *Cell*, vol. 170, no. 1, pp. 17–33, 2017, doi: 10.1016/j.cell.2017.06.009.
- [13] L. Mansi, E. Viel, E. Curtit, J. Medioni, and C. Le Tourneau, "Ciblage de la voie de signalisation RAS pour le traitement des cancers," *Bulletin du Cancer*, vol. 98, no. 9, pp. 1019–1028, 2011, doi: 10.1684/bdc.2011.1380.
- [14] M. P. Di Magliano and C. D. Logsdon, "Roles for KRAS in pancreatic tumor development and progression," *Gastroenterology*, vol. 144, no. 6, pp. 1220–1229, 2013, doi: 10.1053/j.gastro.2013.01.071.
- [15] G. P. Way *et al.*, "Machine learning detects pan-cancer ras pathway activation in the cancer genome atlas," *Cell Reports*, vol. 23, no. 1, pp. 172–180, 2018, doi: 10.1016/j.celrep.2018.03.046.
- [16] T. A. Knijnenburg *et al.*, "Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome atlas," *Cell Reports*, vol. 23, no. 1, pp. 239–254, 2018, doi: 10.1016/j.celrep.2018.03.076.
- [17] J. D. Amaral, J. M. Xavier, C. J. Steer, and C. M. P. Rodrigues, "Targeting the p53 pathway of apoptosis," *Current Pharmaceutical Design*, vol. 16, no. 22, pp. 2493–2503, 2010, doi: 10.2174/138161210791959818.
- [18] P. Monti *et al.*, "Heterogeneity of TP53 mutations and P53 protein residual function in cancer: does it matter?," *Frontiers in Oncology*, vol. 10, 2020, doi: 10.3389/fonc.2020.593383.
- [19] A. Kuşoğlu and Ç. B. Avcı, "Cancer stem cells: a brief review of the current status," *Gene*, vol. 681, pp. 80–85, 2019, doi: 10.1016/j.gene.2018.09.052.
- [20] T. M. Malta *et al.*, "Machine learning identifies stemness features associated with oncogenic dedifferentiation," *Cell*, vol. 173, no. 2, pp. 338–354, 2018, doi: 10.1016/j.cell.2018.03.034.
- [21] R. Thomas and Z. Weihua, "Rethink of EGFR in cancer with its kinase independent function on board," *Frontiers in Oncology*, vol. 9, no. AUG, 2019, doi: 10.3389/fonc.2019.00800.
- [22] R. Qureshi *et al.*, "Machine learning based personalized drug response prediction for lung cancer patients," *Scientific Reports*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-23649-0.
- [23] R. Yang, X. Xiong, H. Wang, and W. Li, "Explainable machine learning model to prediction EGFR mutation in lung cancer," *Frontiers in Oncology*, vol. 12, 2022, doi: 10.3389/fonc.2022.924144.
- [24] M. H. Widyandana *et al.*, "Quercetin as an anticancer candidate for glioblastoma multiforme by targeting AKT1, MMP9, ABCB1, and VEGFA: an in-silico study," *Karbala International Journal of Modern Science*, vol. 9, no. 3, pp. 450–459, 2023, doi: 10.33640/2405-609X.3312.
- [25] J. Eberhardt, D. Santos-Martins, A. F. Tillack, and S. Forli, "AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings," *Journal of Chemical Information and Modeling*, vol. 61, no. 8, pp. 3891–3898, 2021, doi: 10.1021/acs.jcim.1c00203.
- [26] V. V. Kleandrova, L. Scotti, F. J. B. M. Junior, E. Muratov, M. T. Scotti, and A. Speck-Planche, "QSAR modeling for multi-target drug discovery: designing simultaneous inhibitors of proteins in diverse pathogenic parasites," *Frontiers in Chemistry*, vol. 9, 2021, doi: 10.3389/fchem.2021.634663.
- [27] H. Kaneko, "Molecular descriptors, structure generation, and inverse QSAR/QSPR based on SELFIES," *ACS Omega*, vol. 8, no. 24, pp. 21781–21786, 2023, doi: 10.1021/acsomega.3c01332.
- [28] M. R. Iresha *et al.*, "Machine learning model and molecular docking for screening medicinal plants as HIV-1 reverse transcriptase inhibitors," *Karbala International Journal of Modern Science*, vol. 10, no. 1, pp. 79–90, 2024, doi: 10.33640/2405-609X.3341.




- [29] R. Altaf, U. Ilyas, A. Ma, and M. Shi, "Identification and validation of differentially expressed genes for targeted therapy in NSCLC using integrated bioinformatics analysis," *Frontiers in Oncology*, vol. 13, 2023, doi: 10.3389/fonc.2023.1206768.
- [30] 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, 2011, doi: 10.1002/jcc.21707.
- [31] 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, 1988, doi: 10.1021/ci00057a005.
- [32] S. Kim *et al.*, "PubChem in 2021: new data content and improved web interfaces," *Nucleic Acids Research*, vol. 49, no. 1, pp. 1388–1395, 2021, doi: 10.1093/nar/gkaa971.
- [33] Q. Guo *et al.*, "Current treatments for non-small cell lung cancer," *Frontiers in Oncology*, vol. 12, 2022, doi: 10.3389/fonc.2022.945102.
- [34] U. Malapelle *et al.*, "Osimertinib," *Recent Results in Cancer Research*, vol. 211, pp. 257–276, 2018, doi: 10.1007/978-3-319-91442-8_18.
- [35] National Institute of Diabetes and Digestive and Kidney Diseases, "Brigatinib," *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*, Bethesda: National Library of Medicine, 2012.
- [36] National Institute of Diabetes and Digestive and Kidney Diseases, "Alectinib," *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*, Bethesda: National Library of Medicine, 2012.
- [37] A. A. Abdelgalil, H. M. Al-Kahtani, and F. I. Al-Jenoobi, "Erlotinib," *Profiles of Drug Substances, Excipients and Related Methodology*, vol. 45, pp. 93–117, 2020, doi: 10.1016/bs.podrm.2019.10.004.
- [38] National Institute of Child Health and Human, "Ceritinib," *Drugs and Lactation Database (LactMed®)*, Bethesda: National Library of Medicine, 2005.
- [39] E. D. Deeks and G. M. Keating, "Afatinib in advanced NSCLC: a profile of its use," *Drugs and Therapy Perspectives*, vol. 34, no. 3, pp. 89–98, 2018, doi: 10.1007/s40267-018-0482-6.
- [40] Y. Hosomi *et al.*, "Gefitinib alone versus gefitinib plus chemotherapy for non-small-cell lung cancer with mutated epidermal growth factor receptor: NEJ009 study," *Journal of Clinical Oncology*, vol. 38, no. 2, pp. 115–123, 2020, doi: 10.1200/JCO.19.01488.
- [41] E. K. Chung, S. H. Yong, E. H. Lee, E. Y. Kim, Y. S. Chang, and S. H. Lee, "New targeted therapy for non-small cell lung cancer," *Tuberculosis and Respiratory Diseases*, vol. 86, no. 1, pp. 1–13, 2023, doi: 10.4046/trd.2022.0066.
- [42] I. A. Vathiotis, D. Bafaloukos, K. N. Syrigos, and G. Samonis, "Evolving treatment landscape of HER2-mutant non-small cell lung cancer: trastuzumab deruxtecan and beyond," *Cancers*, vol. 15, no. 4, 2023, doi: 10.3390/cancers15041286.
- [43] M. Z. Guo, K. A. Marrone, A. Spira, and S. Rosner, "Adagrasib: a novel inhibitor for KRASG12C-mutated non-small-cell lung cancer," *Future Oncology*, vol. 19, no. 15, pp. 1037–1051, 2023, doi: 10.2217/fon-2022-1106.

BIOGRAPHIES OF AUTHORS






Hamza Hanafi    received his Doctoral degree in Bioinformatics in 2023 and his degree in Computer Science Engineering in 2017, both from the Faculty of Sciences and Technologies at the University Abdelmalek Essaâdi in Tangier, Morocco. He is currently a postdoctoral researcher at the Intelligent Automation and BioMedGenomics Laboratory at the same university. His research interests include computational biology, bioinformatics, and machine learning. He 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.



M'hamed Aït Kbir    is a full professor at the Department of Computer Science, Faculty of Sciences and Technologies 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), and 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 at email: maitkbir@uae.ac.ma.



Badr Dine Rossi Hassani    is a full professor of Biology at the Faculty of Sciences and Technologies of Tangier, Morocco, and Ph.D. managing director at LABIPHABE Laboratory, his research areas interest many disciplines: cancer research, biotechnology, and bioinformatics. He is a member of scientific committees of many international conferences and journals. He can be contacted at email: badrossi@gmail.com.