JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 15 No. 1 (2024) 46-56
https://publisher.uthm.edu.my/ojs/index.php/jscdm

Machine Learning Approach to Predict AXL Kinase Inhibitor Activity for Cancer Drug Discovery Using Bayesian Optimization-XGBoost

Teuku Rizky Noviandy[1,2], Ghalieb Mutig Idroes[2], Irsan Hardi[2]

[1] *Department of Informatics, Faculty of Mathematics and Natural Sciences,*
Universitas Syiah Kuala, Banda Aceh 23111, INDONESIA
[2] *Interdisciplinary Innovation Research Unit,*
Graha Primera Saintifika, Aceh Besar 23771, INDONESIA

*Corresponding Author: trizkynoviandy@gmail.com
DOI: https://doi.org/10.30880/jscdm.2024.05.01.004

Article Info

Received: 1 December 2023
Accepted: 25 April 2024
Available online: 21 June 2024

Keywords

Supervised learning, QSAR, molecular descriptors, ChEMBL, hyperparameter tuning

Abstract

This study aims to predict AXL kinase inhibitors utilizing a Bayesian Optimization-XGBoost machine learning model. A dataset comprising 1074 compounds with IC_{50} values was collected from the ChEMBL database and molecular descriptors for each compound were calculated. The Bayesian Optimization-XGBoost model demonstrated superior performance in predicting AXL kinase inhibitors, achieving an accuracy of 86.24%, precision of 89.52%, recall of 89.52%, and an F1-score of 89.52%, outperforming other models such as LightGBM, Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naïve Bayes. This study underlines the importance of advanced machine learning techniques, particularly Bayesian Optimization-XGBoost, in predicting AXL kinase inhibitors, offering a promising approach for accelerating the early stages of drug discovery. Despite its success, the model's performance depends on the diversity and quality of the training data, and future work should focus on expanding the dataset and validating results with experimental studies. This computational method has the potential to streamline the drug development pipeline and contribute to the discovery of more effective cancer treatments.

1. Introduction

Cancer continues to be a serious worldwide health issue, characterized by unregulated cell proliferation and the ability to spread to other parts of the body, therefore providing a substantial risk to human health and welfare [1]. In recent years, advancements in molecular biology and targeted therapies have opened new avenues for cancer treatment [2,3]. Among the promising targets, AXL kinase has gained attention as a significant contributor to the advancement and spread of cancer [4]. This kinase belongs to the TAM (Tyro3, AXL, Mer) family and is implicated in various malignancies, including but not limited to breast, lung, and pancreatic cancers [5]. Given its crucial involvement in cancer biology, inhibiting AXL kinase activity has emerged as a promising strategy for developing novel anticancer therapeutics [6,7].

The AXL receptor tyrosine kinase, which belongs to the TAM family, has a vital function in controlling several cellular processes such as cell survival, proliferation, and migration [8]. AXL is often overexpressed in cancer cells, contributing to their ability to evade apoptosis, promote angiogenesis, and facilitate metastasis [9]. Its association with epithelial-mesenchymal transition (EMT) and resilience to traditional therapies underscores its significance as a potential target for cancer drug development. Consequently, there is a growing interest in elucidating the

This is an open access article under the CC BY-NC-SA 4.0 license.

molecular mechanisms underlying AXL kinase activity and developing specific inhibitors to modulate its function for therapeutic purposes.

In recent years, the integration of machine learning techniques in drug discovery has revolutionized the process of identifying potential therapeutic agents [10–13]. Machine learning has been a prominent method in Quantitative Structure-Activity Relationship (QSAR) modeling. It is used to predict the biological activity of substances by analyzing their structural characteristics [14,15]. QSAR explores the intricate relationship between a compound's structure and its biological activity, facilitating the prediction of how structural modifications may influence efficacy [16]. Machine learning algorithms offer the advantage of efficiently analyzing complex relationships within large datasets, providing insights that can significantly expedite the drug development pipeline [17–19].

Among the various machine learning methods employed in drug discovery, XGBoost has emerged as a popular and powerful algorithm [20]. XGBoost, an ensemble learning technique, is well-suited for handling diverse data types and achieving high accuracy [21,22]. However, to harness its full potential, optimizing hyperparameters is crucial. Bayesian optimization, a probabilistic model-based optimization approach, has proven effective in fine-tuning the hyperparameters of complex machine learning models. Its ability to guide the search for optimal parameter configurations makes Bayesian optimization a valuable companion to enhance the performance of XGBoost in drug discovery.

In this paper, we introduce an approach to predict AXL kinase inhibitors activity through the utilization of the Bayesian Optimization-XGBoost approach. Our contribution involves creating our own dataset by collecting compounds data from ChEMBL and calculating molecular descriptors of each compound to be used as features. Through a comprehensive comparative analysis against five other prevalent machine learning models used in AXL kinase inhibitor activity prediction, we establish the superior performance of our Bayesian Optimization-XGBoost model. The results confirm that our approach surpasses other models and highlights the significance of utilizing advanced machine learning techniques. This emphasizes the potential of Bayesian Optimization-XGBoost as a powerful tool for predicting AXL kinase inhibitors activity in drug discovery applications.

2. Related Works

Machine learning has become a potent tool in the field of cancer medication discovery in recent years, providing researchers with new approaches to find and create efficient treatments. By leveraging the predictive capabilities of machine learning algorithms, scientists can now predict and prioritize compounds based on their likelihood of inhibiting cancer growth and progression. This approach has the potential to significantly accelerate the drug discovery process, allowing researchers to focus their efforts on the most promising candidates. Several notable studies have already demonstrated the effectiveness of this strategy, showcasing impressive results that highlight the immense potential of machine learning in the fight against cancer.

One notable work explored the use of a random forest method to predict androgen binding, agonistic, and antagonistic action in cancer. Their prediction models demonstrated commendable accuracy, with agonists predicted at an 80% rate, while the respective metrics for antagonists and binders were 72% and 78% [23]. Further study investigated the role of the androgen receptor and utilized QSAR and machine learning techniques, namely the Extreme Learning Machine algorithm. The model demonstrated a noteworthy prediction r-squared value of 0.737, suggesting its efficacy in forecasting the antiproliferative potential of these chemicals [24]. Additionally, a study conducted on VEGFR-2 kinase inhibitors, which are medications approved for clinical use that specifically target the growth of blood vessels in cancer, showcased the strength and reliability of the 2D-SAR model. The K-Nearest Neighbors (KNN) approach demonstrated a prediction accuracy of 82.4% for the training set and 80.1% for the test set. This highlights the reliable predictive capability of the model in discovering effective VEGFR-2 kinase inhibitors [25].

Another study applied random forest techniques to predict inhibitors of Estrogen Receptor Alpha in breast cancer drug discovery. Their model achieved an accuracy of 0.745, showcasing its capability to discern potential inhibitors effectively. This contribution underscores the versatility of machine learning in predicting inhibitors for specific cancer-related targets [26]. Furthermore, the challenge of predicting AXL kinase inhibitors was tackled through gradient boosting techniques, analyzing a dataset of 527 compounds. The resulting model boasted an accuracy of 0.85. Nonetheless, the discrepancy between accuracy and the F1-score hinted at a potential class imbalance issue [13]. While the accuracy suggests proficient prediction for one class, the F1-score reveals room for improvement in achieving balanced performance across multiple classes. This underscores the necessity for further research to enhance the model's capability in accurately predicting AXL Kinase Inhibitor Activity for a diverse range of compounds. The pursuit of a more nuanced and balanced model performance is crucial for its practical applicability in the complex landscape of cancer.

3. Methods

A concise overview of the framework in this study is depicted in Fig. 1. Initially, dataset preparation is undertaken, followed by the execution of machine learning modeling to predict AXL kinase inhibitor activity.

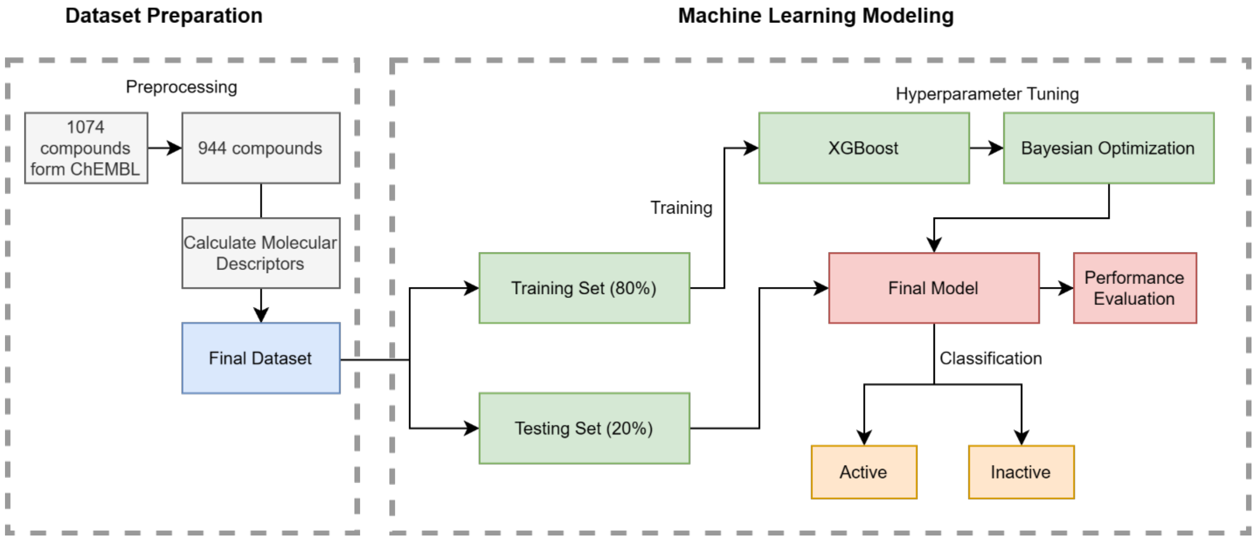

Fig. 1. *Framework of Bayesian Optimization-XGBoost for AXL Kinase inhibitor prediction*

3.1 Data Collection and Preprocessing

A total of 1074 compounds and their corresponding IC_{50} values was obtained from the ChEMBL database [27]. The IC_{50} values represent the half-maximal inhibitory concentration, quantifying the potency of compounds in inhibiting AXL kinase activity. A systematic pre-processing protocol was implemented to ensure data integrity and reliability. Duplicate entries were removed, and any instances with missing values were addressed. Following these steps, a total of 944 compounds remained in the dataset. Subsequently, the dataset was categorized into two classes based on IC_{50} values. Compounds with IC_{50} values less than 1000 were designated as the "active" class, indicating higher inhibitory activity, while those with IC_{50} values greater than or equal to 1000 were classified as the "inactive" class [28,29]. This resulted in 578 active compounds and 366 inactive compounds.

Following the pre-processing procedures, the dataset was partitioned into separate training and testing sets to ease the training and evaluation of the model. The compounds were divided into two subsets: 80% of the data (755 compounds) was assigned to the training set, while the remaining 20% (189 compounds) formed the testing set. This division guarantees that the model is trained on a dataset of appropriate size and assessed on a distinct, unseen subset to evaluate its ability to generalize [30].

3.2 Molecular Descriptors

Molecular descriptors are numerical representations of chemical compounds that encode information about their physicochemical, structural, and electronic properties [31,32]. These descriptors serve as a crucial bridge between chemical structures and their corresponding biological activities, aiding in the quantitative analysis and prediction of molecular behavior. In this study, molecular descriptors were calculated using the Mordred Python library [33].

For each of the 944 compounds in our dataset, Mordred generated a set of 1074 molecular descriptors. These descriptors collectively provide each compound's detailed and multidimensional profile, capturing essential information that contributes to the compounds' AXL kinase inhibitory activity. The extensive nature of these descriptors allows for a thorough exploration of the structural and chemical properties influencing the biological function of the compounds.

3.3 XGBoost Model

XGBoost is a member of the ensemble learning family and is a machine learning algorithm. XGBoost combines the strengths of both boosting and regularization to create a robust and accurate predictive model. Boosting involves training a sequence of weak learners, typically decision trees, where each subsequent tree corrects the errors of the previous ones [34]. Regularization is introduced through shrinkage (or learning rate) and feature importance regularization, preventing overfitting and enhancing the model's generalization capabilities [35].

The success of XGBoost can be attributed to its ability to handle complex relationships within the data and provide feature importance scores. This study employs XGBoost to predict AXL kinase inhibitor activity based on molecular descriptors. The general equation for the prediction made by XGBoost can be expressed as follows:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{1}$$

Where $\hat{y}_i$ is the predicted output for the i-th data point, K is the number of weak learners (trees), and $f_k(x_i)$ represents the output of the k-th weak learner for the input x_i.

3.4 Bayesian Optimization

XGBoost requires effective tuning of hyperparameters to unleash its full potential. Bayesian optimization is an efficient method for this purpose, surpassing traditional approaches like random or grid search [36]. Bayesian optimization is especially beneficial in situations when the assessment of the objective function (performance measure) requires a significant amount of computer resources or takes a long time to complete [37,38].

Bayesian optimization is a method of optimizing a function that uses probabilistic surrogate models to estimate the objective function. It is a sequential process that iteratively improves the model's approximation [39]. The core concept is to strategically choose the subsequent set of hyperparameters to assess, using on the surrogate model's predictions and measures of uncertainty. This iterative procedure progressively improves the model and approaches the best combination of hyperparameters [40].

The process initiates with defining a prior distribution over the function f, to be optimized, embodying our initial beliefs before observing any data. This prior, represented by a Gaussian Process in Bayesian optimization, encodes assumptions about the smoothness and variability of f. Gaussian Processes are favored for their flexibility and capacity to model the prior over the objective function with a set of kernel parameters, which dictate the function's smoothness and correlation structure [41].

Upon each iteration, Bayesian optimization combines the prior with the likelihood of observed data (objective function evaluations) to form a posterior distribution over f. This posterior reflects the updated beliefs about the function’s behavior and guides the selection of new hyperparameter configurations by predicting their expected performance. The balance between exploration and exploitation is maintained by an acquisition function, which leverages the posterior to identify points with the highest utility [42].

The posterior distribution is derived in this study by applying Bayes' theorem to combine the likelihood function and the prior distribution. Bayes' theorem states that the posterior distribution of a parameter, given the observed data, is directly proportional to the product of the probability function of the data given the parameter and the prior distribution of the parameter [43]. In Bayesian optimization, the surrogate model's posterior distribution over the objective function is updated iteratively as new data points (objective function evaluations) are observed. Initially, the prior distribution is established based on prior knowledge or assumptions about the objective function. As more data becomes available, the posterior distribution becomes more informed, leading to improved predictions and uncertainty estimates. This iterative process of updating the posterior distribution forms the foundation of Bayesian optimization's efficiency in searching for the optimal set of hyperparameters. [44]

The acquisition function, often denoted as $\alpha(x)$ is a key component of Bayesian optimization. It quantifies the utility of evaluating a specific set of hyperparameters and is used to decide the next configuration to explore. The acquisition function used is the Expected Improvement (EI), which balances the potential improvement in objective function value with uncertainty, as shown in Equation 2 [45].

$$EI(x) = \mathbb{E}[\max(f(x_{best}) - f(x), 0)] \tag{2}$$

Where x is the set of hyperparameters to be evaluated, $f(x)$ is the predicted objective function value at x from the surrogate model, x_{best} is the current best-known set of hyperparameters, and $\mathbb{E}$ denotes the expectation.

The hyperparameter space defined for XGBoost tuning in this study is presented in Table 1. The ranges specified for each hyperparameter guide the Bayesian optimization process, allowing the algorithm to explore and optimize the XGBoost model for the prediction of AXL kinase inhibitor activity. The maximum depth of every tree is specified by the'max_depth' option. 'learning_rate' regulates the optimization process's step size. 'n_estimators' indicates how many rounds of boosting there will be. The value 'min_child_weight' determines the lowest total instance weight necessary for a child. The fraction of training data used in each boosting round is determined by the variable "subsample." It is specified by 'colsample_bytree' how many features are used in each boosting cycle. The minimal loss reduction required to create a second division on a leaf node is controlled by "gamma."

Table 1 *Hyperparameter space for the Bayesian optimization*

Hyperparameter	Range
max_depth	(5, 25)
learning_rate	(0.01, 0.5)
n_estimators	(50, 150)
min_child_weight	(1, 10)
subsample	(0.5, 1)
colsample_bytree	(0.5, 1)
gamma	(0, 1)

3.5 Performance Evaluation

The Bayesian Optimization-XGBoost model's ability to predict AXL kinase inhibitor action was extensively evaluated using key metrics including accuracy, precision, recall, and F1-score. Accuracy is a metric that quantifies the overall correctness of a model's predictions. Precision is a statistical measure that quantifies the proportion of accurately anticipated positive observations out of the total number of predicted positives. Recall assesses the model's capacity to accurately identify all true positive instances. The F1-score is a metric that considers both precision and recall in a balanced manner. The formulas for accuracy, precision, recall, and F1-score are provided in formulas 3, 4, 5, and 6, respectively [46,47].

$$Accuracy = \frac{TP + FN}{FP + FN + TP + TN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{FN + TP} \tag{5}$$

$$F1 - score = 2\frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

TP (True Positives) are instances correctly identified as belonging to the positive class, while FP (False Positives) are instances incorrectly classified as positive. TN (True Negatives) are instances correctly classified as negative, and FN (False Negatives) are instances incorrectly labeled as negative [48].

In addition to evaluating the XGBoost model, a further analysis was conducted by comparing its performance against five other models: LightGBM, Logistic Regression, KNN, Support Vector Machine (SVM), and Naïve Bayes. This comparative assessment aimed to provide insights into the relative strengths and weaknesses of the XGBoost model for AXL kinase inhibitor activity prediction for cancer drug discovery.

4. Results

This study has successfully trained and optimized an XGBoost model using Bayesian optimization. The Bayesian optimization process for hyperparameter tuning in LightGBM resulted in selecting the most optimal configuration, enhancing the model's performance in classifying AXL kinase inhibitor activity. The best hyperparameter set, with a corresponding target value of 0.8624, includes values such as 'colsample_bytree': 0.6239, 'gamma': 0.5612, 'learning_rate': 0.4691, 'max_depth': 5, 'min_child_weight': 8, 'n_estimators': 95, and 'subsample': 0.7279.

The specified 'colsample_bytree' value of 0.6239 indicates the fraction of features utilized during each boosting round, optimizing the model's ability to capture diverse information. The 'gamma' value of 0.5612 regulates the minimum loss reduction required for further partitioning, contributing to efficient tree pruning and preventing overfitting. A 'learning_rate' of 0.4691 controls the step size during optimization, influencing the model's adaptability and convergence speed. With a 'max_depth' of 5, the model achieves an optimal balance between capturing complex relationships and preventing overfitting. The 'min_child_weight' value of 8 sets the minimum sum of instance weight required in a child, influencing the model's sensitivity to small variations. 'N_estimators' at 95 defines the number of boosting rounds, balancing computational efficiency and model accuracy. Lastly, 'subsample' at 0.7279 determines the fraction of training samples used in each boosting round, contributing to robustness and preventing overfitting.

The performance of the AXL kinase inhibitor activity classification models is presented in Table 2. XGBoost demonstrates superior performance across all metrics, attaining an accuracy of 86.24%, precision of 89.52%, recall of 89.52%, and an F1-score of 89.52%. However, it's notable that LightGBM surpasses XGBoost in recall, achieving an impressive score of 95.97%. Despite LightGBM's higher recall, XGBoost still stands out for its remarkable precision, indicating a low rate of false positives and a strong ability to correctly identify compounds with true inhibitory activity. Additionally, XGBoost demonstrates balanced performance reflected in its F1-score, highlighting its proficiency in capturing both true positive and true negative instances. This balance indicates the model's ability to effectively identify compounds with AXL kinase inhibitory potential while minimizing the risk of false negatives.

In terms of run time, the models in Table 2 exhibit relatively similar performance, with XGBoost requiring just 0.11 seconds for execution. While this run time is marginally longer than that of some models like KNN and Naïve Bayes, it remains within an acceptable range and is not significantly different. This slight difference in run time is unlikely to impact the practical usability of XGBoost, especially considering its outstanding performance in accurately classifying AXL kinase inhibitor activity. Thus, despite minor variations in run time, the overall efficiency of XGBoost remains compelling, further supporting its preference as the primary choice for classification tasks.

Table 2 *Performance metrics of AXL kinase inhibitor activity classification models*

Model	Run Time (Seconds)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Bayesian Optimization-XGBoost	0.11	86.24	89.52	89.52	89.52
LightGBM	0.12	84.13	82.64	95.97	88.81
Logistic Regression	0.11	83.60	88.43	86.29	87.35
KNN	0.01	81.48	86.18	85.48	85.83
SVM	0.15	84.66	89.26	87.10	88.16
Naïve Bayes	0.01	75.66	79.55	84.68	82.03

The ROC curve analysis revealed a comparative assessment of various machine learning models for classifying AXL kinase inhibitor activity, a crucial parameter in cancer drug discovery (Fig. 2). The XGBoost model demonstrated the highest discriminative performance, boasting an AUC of 0.908. Close behind were the SVM and Logistic Regression models, with AUCs of 0.907 and 0.899, respectively. These models excelled in distinguishing between active and inactive compounds. The LightGBM model achieved an AUC of 0.867, while the KNN model followed closely with an AUC of 0.864, indicating moderate performance. However, the Naïve Bayes model fell short with an AUC of 0.813, suggesting lower specificity and sensitivity in classification tasks within this study.

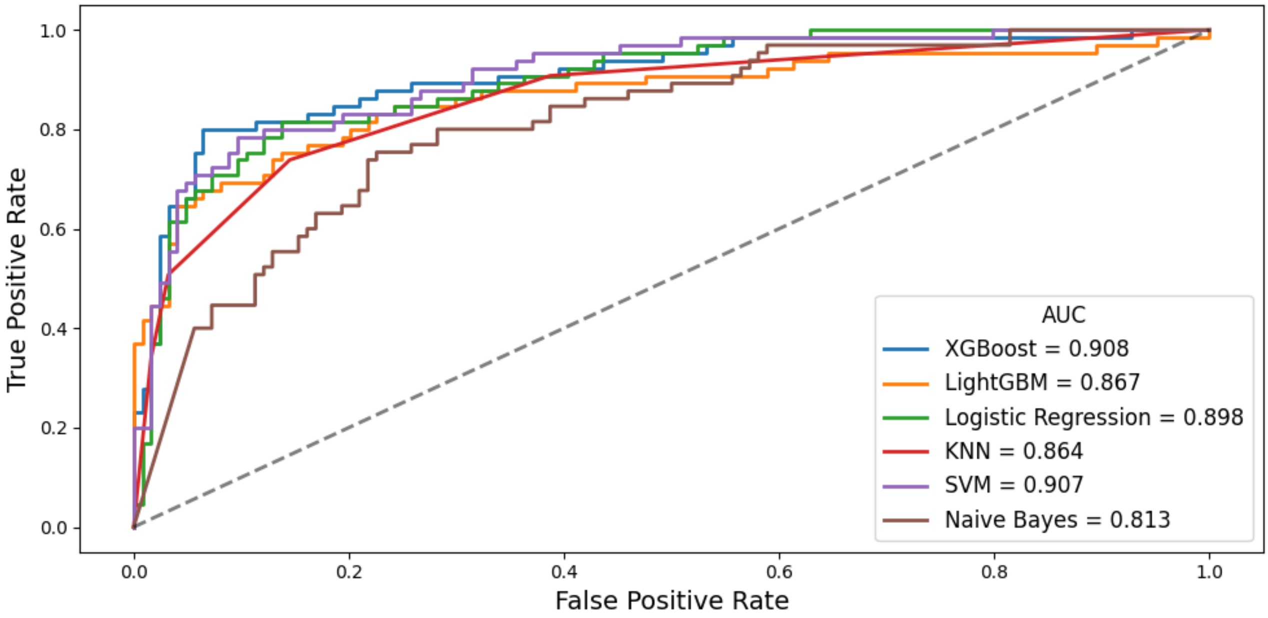

Fig. 2 *ROC curve of the machine learning models*

In the XGBoost model employed for classifying AXL kinase inhibitor activity, the top five molecular descriptors, as visualized in Fig. 3, provided key insights into the structural attributes that were most predictive of inhibitor efficacy. The most significant descriptor, ATSCi, represents a 2D autocorrelation descriptor weighted by ionization potential, capturing the influence of electron distribution on molecular interaction. This descriptor captures how electron distribution influences molecular interactions, which is crucial in understanding how inhibitors interact with the AXL kinase active site. EState_VSA9, the second most important feature, reflects the electronic state for a specific van der Waals surface area, indicative of the compound's potential reactivity and interaction surface. Van der Waals interactions are essential in molecular recognition and binding, and this descriptor helps gauge the compound's potential reactivity and interaction surface, which is vital for effective kinase inhibition. The AATSC1dv descriptor, slightly less influential, associated with a 3D autocorrelation descriptor influenced by valence electrons, offering spatial and electronic information pertinent to binding affinity. The nHBacc descriptor quantifies the number of hydrogen bond acceptors, a crucial factor in the bioactivity of kinase inhibitors, as these interactions often stabilize the inhibitor-kinase complex. Lastly, nBase represents the count of basic atoms, providing insight into the compound's ability to engage in protonation or charge-charge interactions, which can be critical for biological activity. Together, these descriptors form a comprehensive profile that the model uses to distinguish between active and inactive AXL kinase inhibitors.

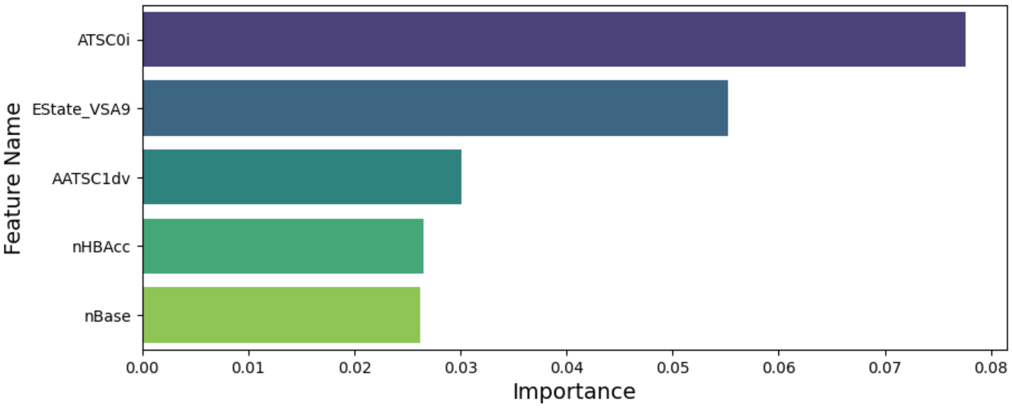

Fig. 3 *Feature importance of the XGBoost model*

To assess the applicability domain of our XGBoost model, we visualized the Principal Component Analysis (PCA) plot in Fig. 4. In this figure, PC-1 and PC-2 represented the first two principal components, which accounted for the majority of variation in the molecular descriptor dataset used for model training. The results showed that PC-1 explained 18.73% of the total variance and PC-2 accounted for 13.60% of the total variance. While 32.33% might have seemed relatively low, given the high number of molecular descriptors used, this percentage was quite substantial, indicating a significant portion of the dataset's variability was captured within these two principal components.

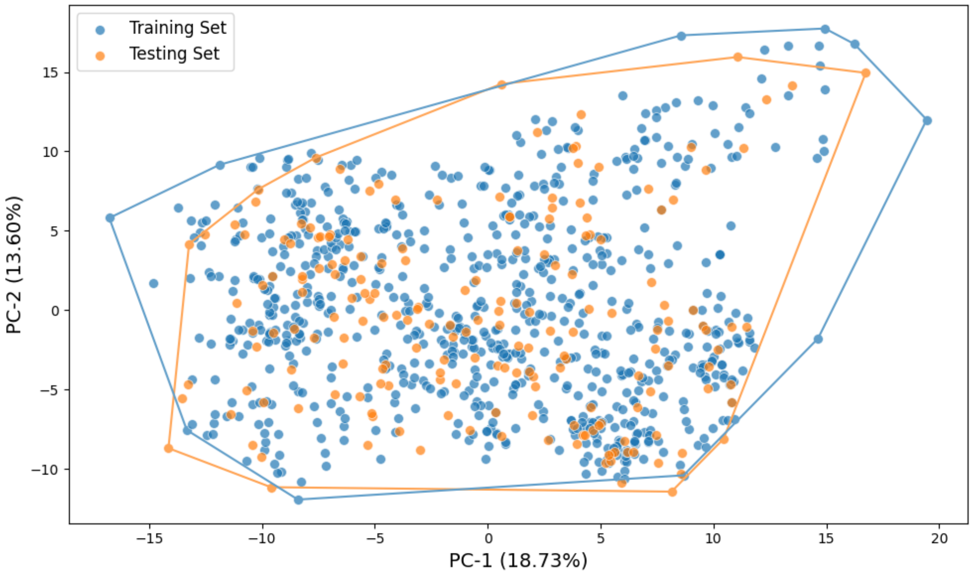

Fig. 4 *Applicability domain of the XGBoost model*

The distribution of points, with the training set represented by blue points and the testing set by orange points, suggests a strong overlap between the two datasets. This overlap indicates that the testing set is a good representation of the chemical space covered by the training data. The convex hulls drawn around each dataset delineate the outer boundaries within which the model's predictions are deemed reliable. The substantial overlap between the training and testing sets within these boundaries suggests that the model generalizes well and has a broad applicability domain. This overlap implies that the model is likely to make accurate predictions for new compounds as long as they fall within this defined domain. However, compounds lying outside of these boundaries may be beyond the model's applicability domain, and predictions for such compounds would be less reliable or considered extrapolations.

5. Discussion

The present study introduces a sophisticated computational approach to cancer drug discovery by leveraging an XGBoost model enhanced with Bayesian Optimization for the classification of AXL kinase inhibitor activity. By employing a comprehensive set of molecular descriptors, the model predicts the inhibitory potential of chemical compounds with high accuracy, holding significant implications for expediting the early stages of drug discovery and development. The identification of the most predictive features of AXL kinase inhibitors not only optimizes the search for novel cancer therapeutics but also contributes to a deeper understanding of the molecular mechanisms underlying inhibitor efficacy.

The implications of this study extend beyond the computational domain and into the practical aspects of pharmaceutical research. The developed predictive model can serve as a powerful preliminary screening tool, enabling researchers to prioritize compounds for synthesis and biological testing. By focusing resources on the most promising candidates, this approach has the potential to significantly reduce the time and cost associated with the drug development pipeline. The model's high classification accuracy can streamline the subsequent lab-based validation processes, ultimately leading to the discovery of more targeted and effective cancer treatments.

Moreover, the feature importance analysis conducted in this study provides valuable insights into the molecular properties that contribute to AXL kinase inhibitor activity. By identifying the key descriptors that influence inhibitor efficacy, this research lays the groundwork for the rational design of novel compounds with enhanced therapeutic potential. The knowledge gained from this analysis can guide medicinal chemists in the optimization of lead compounds, enabling the development of more potent and selective AXL kinase inhibitors.

Nevertheless, the study had certain limitations. The effectiveness of the predictive model is inherently dependent on the quality and scope of the training data. If the training dataset lacks diversity or is biased towards certain chemical spaces, the model's applicability domain may be limited, potentially overlooking viable inhibitor candidates. To address this issue, future studies should aim to expand the chemical space of the training data by incorporating a broader range of molecular descriptors and inhibitor classes. This expansion will enhance the model's generalizability and increase its potential to identify novel AXL kinase inhibitors from diverse chemical backgrounds.

Furthermore, it is crucial to recognize that the in-silico predictions generated by the model need to be substantiated through in vitro and in vivo studies. While computational methods provide valuable insights and predictions, the biological relevance of these findings must be validated experimentally. Collaborations between computational chemists and experimental biologists will be essential to bridge the gap between in silico predictions and real-world cancer therapies. By integrating computational findings with experimental validation, researchers can gain a more comprehensive understanding of the mechanisms of action and potential side effects of the predicted inhibitors.

6. Conclusion

This study demonstrates the potential of Bayesian Optimization-XGBoost to classify AXL Kinase inhibitor activity, which could expedite and refine the search for new cancer treatments. The Bayesian Optimization-XGBoost model obtained an accuracy of 86.24%, precision of 89.52%, recall of 89.52%, and an F1-score of 89.52%, outperforming other models such as LightGBM, Logistic Regression, KNN, SVM, and Naïve Bayes. Despite the promising results, the reliability of these predictions is inherently tied to the dataset's comprehensiveness and the model's applicability domain. Future endeavors should focus on enhancing the dataset diversity, corroborating computational predictions with experimental data, and expanding the model's utility through the inclusion of additional biological parameters. Through such improvements, this computational approach holds the promise of becoming an invaluable asset in the ongoing quest for effective cancer therapeutics.

Appendix A: Source Code

The source code utilized for this study is available at: https://github.com/trizkynoviandy/ML-AXL-Kinase

Acknowledgement

The authors would like to thank the Heca Sentra Analitika for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Teuku Rizky Noviandy, Ghalieb Mutig Idroes;* ***data collection:*** *Teuku Rizky Noviandy;* ***analysis and interpretation of results:*** *Teuku Rizky Noviandy, Irsan Hardi;* ***draft manuscript preparation:*** *Teuku Rizky Noviandy, Ghalieb Mutig Idroes, Irsan Hardi. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Myers, S. H., Brunton, V. G., & Unciti-Broceta, A. (2016). AXL Inhibitors in Cancer: A Medicinal Chemistry Perspective. *Journal of Medicinal Chemistry*, *59*(8), 3593–3608. https://doi.org/10.1021/acs.jmedchem.5b01273

[2] Lee, Y. T., Tan, Y. J., & Oon, C. E. (2018). Molecular targeted therapy: Treating cancer with specificity. *European Journal of Pharmacology*, *834*, 188–196. https://doi.org/10.1016/j.ejphar.2018.07.034

[3] Huang, M., Shen, A., Ding, J., & Geng, M. (2014). Molecularly targeted cancer therapy: some lessons from the past decade. *Trends in Pharmacological Sciences*, *35*(1), 41–50. https://doi.org/10.1016/j.tips.2013.11.004

[4] Goyette, M.-A., & Côté, J.-F. (2022). AXL Receptor Tyrosine Kinase as a Promising Therapeutic Target Directing Multiple Aspects of Cancer Progression and Metastasis. *Cancers*, *14*(3), 466. https://doi.org/10.3390/cancers14030466

[5] Malvankar, C., & Kumar, D. (2022). AXL kinase inhibitors- A prospective model for medicinal chemistry strategies in anticancer drug discovery. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1877*(5), 188786. https://doi.org/10.1016/j.bbcan.2022.188786

[6] Li, P., Niu, Y., Li, S., Zu, X., Xiao, M., Yin, L., Feng, J., He, J., & Shen, Y. (2022). Identification of an AXL kinase inhibitor in triple-negative breast cancer by structure-based virtual screening and bioactivity test. *Chemical Biology & Drug Design*, *99*(2), 222–232. https://doi.org/10.1111/cbdd.13977

[7] Ozyurt, R., & Ozpolat, B. (2023). Therapeutic Landscape of AXL Receptor Kinase in Triple-Negative Breast Cancer. *Molecular Cancer Therapeutics*, *22*(7), 818–832. https://doi.org/10.1158/1535-7163.MCT-22-0617

[8] Engelsen, A. S. T., Lotsberg, M. L., Abou Khouzam, R., Thiery, J.-P., Lorens, J. B., Chouaib, S., & Terry, S. (2022). Dissecting the Role of AXL in Cancer Immune Escape and Resistance to Immune Checkpoint Inhibition. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.869676

[9] Li, Y., Ye, X., Tan, C., Hongo, J.-A., Zha, J., Liu, J., Kallop, D., Ludlam, M. J. C., & Pei, L. (2009). Axl as a potential therapeutic target in cancer: role of Axl in tumor growth, metastasis and angiogenesis. *Oncogene*, *28*(39), 3442–3455. https://doi.org/10.1038/onc.2009.212

[10] Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., & Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery. *Malacca Pharmaceutics*, *1*(2), 48–54. https://doi.org/10.60084/mp.v1i2.60

[11] Kurniawan, I., Rosalinda, M., & Ikhsan, N. (2020). Implementation of ensemble methods on QSAR Study of NS3 inhibitor activity as anti-dengue agent. *SAR and QSAR in Environmental Research*, *31*(6), 477–492. https://doi.org/10.1080/1062936X.2020.1773534

[12] Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., & Idroes, R. (2023). QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer's Disease Using Ensemble Machine Learning Algorithms. *Heca Journal of Applied Sciences*, *1*(1), 1–7. https://doi.org/10.60084/hjas.v1i1.12

[13] Khan, M. B., Shahrior, R., Asha, R. T., & Saha, P. S. (2021). Predicting AXL Inhibition of Chemicals using Molecular Descriptors and Machine Learning Methods. *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*, 1–6. https://doi.org/10.1109/EICT54103.2021.9733504

[14] Hesping, E., Chua, M. J., Pflieger, M., Qian, Y., Dong, L., Bachu, P., Liu, L., Kurz, T., Fisher, G. M., Skinner-Adams, T. S., Reid, R. C., Fairlie, D. P., Andrews, K. T., & Gorse, A.-D. J. P. (2022). QSAR Classification Models for Prediction of Hydroxamate Histone Deacetylase Inhibitor Activity against Malaria Parasites. *ACS Infectious Diseases*, *8*(1), 106–117. https://doi.org/10.1021/acsinfecdis.1c00355

[15] Noviandy, T. R., Maulana, A., Idroes, G. M., Emran, T. B., Tallei, T. E., Helwani, Z., & Idroes, R. (2023). Ensemble Machine Learning Approach for Quantitative Structure Activity Relationship Based Drug Discovery: A Review. *Infolitika Journal of Data Science*, *1*(1), 32–41. https://doi.org/10.60084/ijds.v1i1.91

[16] Noviandy, T. R., Nisa, K., Idroes, G. M., Hardi, I., & Sasmita, N. R. (2024). Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer's Drug Discovery with LightGBM. *Journal of Computing Theories and Applications*, *2*(2), 138–147. https://doi.org/10.62411/jcta.10129

[17] Idroes, R., Noviandy, T., Maulana, A., Suhendra, R., Sasmita, N., Muslem, M., Idroes, G. M., Kemala, P., & Irvanizam, I. (2021). Application of Genetic Algorithm-Multiple Linear Regression and Artificial Neural Network Determinations for Prediction of Kovats Retention Index. *International Review on Modelling and Simulations (IREMOS)*, *14*(2), 137.

[18] Agustia, M., Noviandy, T. R., Maulana, A., Suhendra, R., Muslem, M., Sasmita, N. R., Idroes, G. M., Rahimah, S., Afidh, R. P. F., Subianto, M., Irvanizam, I., & Idroes, R. (2022). Application of Fuzzy Support Vector Regression to Predict the Kovats Retention Indices of Flavors and Fragrances. *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)*, 13–18. https://doi.org/10.1109/ICELTICs56128.2022.9932124

[19] Noviandy, T. R., Maulana, A., Sasmita, N. R., Suhendra, R., Irvanizam, I., Muslem, M., Idroes, G. M., Yusuf, M., Sofyan, H., Abidin, T. F., & Idroes, R. (2022). The Prediction of Kovats Retention Indices of Essential Oils at Gas Chromatography Using Genetic Algorithm-Multiple Linear Regression and Support Vector Regression. *Journal of Engineering Science and Technology*, *17*(1), 306–326.

[20] Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–794.

[21] Maulana, A., Noviandy, T. R., Sasmita, N. R., Paristiowati, M., Suhendra, R., Yandri, E., Satrio, J., & Idroes, R. (2023). Optimizing University Admissions: A Machine Learning Perspective. *Journal of Educational Management and Learning*, *1*(1), 1–7. https://doi.org/10.60084/jeml.v1i1.46

[22] Maulana, A., Faisal, F. R., Noviandy, T. R., Rizkia, T., Idroes, G. M., Tallei, T. E., El-Shazly, M., & Idroes, R. (2023). Machine Learning Approach for Diabetes Detection Using Fine-Tuned XGBoost Algorithm. *Infolitika Journal of Data Science*, *1*(1), 1–7. https://doi.org/10.60084/ijds.v1i1.72

[23] Piir, G., Sild, S., & Maran, U. (2021). Binary and multi-class classification for androgen receptor agonists, antagonists and binders. *Chemosphere*, *262*, 128313. https://doi.org/10.1016/j.chemosphere.2020.128313

[24] Oyeneyin, O. E., Obadawo, B. S., Metibemu, D. S., Owolabi, T. O., Olanrewaju, A. A., Orimoloye, S. M., Ipinloju, N., & Olusayo, O. (2022). An exploration of the antiproliferative potential of chalcones and dihydropyrazole derivatives in prostate cancer via androgen receptor: combined QSAR, machine learning, and molecular docking techniques. *Physical Chemistry Research*, *10*(2), 211–223.

[25] Ding, H., Xing, F., Zou, L., & Zhao, L. (2024). QSAR analysis of VEGFR-2 inhibitors based on machine learning, Topomer CoMFA and molecule docking. *BMC Chemistry*, *18*(1), 59. https://doi.org/10.1186/s13065-024-01165-8

[26] Liu, J., Zhou, Z., Kong, S., & Ma, Z. (2022). Application of random forest based on semi-automatic parameter adjustment for optimization of anti-breast cancer drugs. *Frontiers in Oncology*, *12*. https://doi.org/10.3389/fonc.2022.956705

[27] Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777

[28] Noviandy, T. R., Maulana, A., Idroes, G. M., Irvanizam, I., Subianto, M., & Idroes, R. (2023). QSAR-Based Stacked Ensemble Classifier for Hepatitis C NS5B Inhibitor Prediction. *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, 220–225. https://doi.org/10.1109/COSITE60233.2023.10250039

[29] Simeon, S., Anuwongcharoen, N., Shoombuatong, W., Malik, A. A., Prachayasittikul, V., Wikberg, J. E. S., & Nantasenamat, C. (2016). Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. *PeerJ*, *4*, e2322.

[30] Noviandy, T. R., Nainggolan, S. I., Raihan, R., Firmansyah, I., & Idroes, R. (2023). Maternal Health Risk Detection Using Light Gradient Boosting Machine Approach. *Infolitika Journal of Data Science*, *1*(2), 48–55. https://doi.org/10.60084/ijds.v1i2.123

[31] Mauri, A., Consonni, V., & Todeschini, R. (2017). Molecular Descriptors. In *Handbook of Computational Chemistry* (pp. 2065–2093). Springer International Publishing. https://doi.org/10.1007/978-3-319-27282-5_51

[32] Maulydia, N. B., Khairan, K., & Noviandy, T. R. (2023). Prediction of Pharmacokinetic Parameters from Ethanolic Extract Mane Leaves (Vitex pinnata L.) in Geothermal Manifestation of Seulawah Agam Ie-Seu'um, Aceh. *Malacca Pharmaceutics*, *1*(1), 16–21. https://doi.org/10.60084/mp.v1i1.33

[33] Moriwaki, H., Tian, Y. S., Kawashita, N., & Takagi, T. (2018). Mordred: A molecular descriptor calculator. *Journal of Cheminformatics*, *10*(1), 1–14. https://doi.org/10.1186/s13321-018-0258-y

[34] Suhendra, R., Suryadi, S., Husdayanti, N., Maulana, A., Noviandy, T. R., Sasmita, N. R., Subianto, M., Earlia, N., Niode, N. J., & Idroes, R. (2023). Evaluation of Gradient Boosted Classifier in Atopic Dermatitis Severity Score Classification. *Heca Journal of Applied Sciences*, *1*(2), 54–61. https://doi.org/10.60084/hjas.v1i2.85

[35] Maulana, A., Noviandy, T. R., Suhendra, R., Earlia, N., Sofyan, H., Subianto, M., & Idroes, R. (2023). Performance Analysis and Feature Extraction for Classifying the Severity of Atopic Dermatitis Diseases. *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, 226–231. https://doi.org/10.1109/COSITE60233.2023.10249760

[36] Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1), 26–40.

[37] Brochu, E., Cora, V. M., & De Freitas, N. (2010). A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *ArXiv Preprint ArXiv:1012.2599*.

[38] Idroes, R., Maulana, A., Noviandy, T. R., Suhendra, R., Sasmita, N. R., Lala, A., & Irvanizam. (2020). A Genetic Algorithm to Determine Research Consultation Schedules in Campus Environment. *IOP Conference Series: Materials Science and Engineering*, *796*, 012033. https://doi.org/10.1088/1757-899X/796/1/012033

[39] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, *104*(1), 148–175. https://doi.org/10.1109/JPROC.2015.2494218

[40] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, *25*.

[41] Seeger, M. (2004). Gaussian Processes For Machine Learning. *International Journal of Neural Systems*, *14*(02), 69–106. https://doi.org/10.1142/S0129065704001899

[42] Zhang, S., Yang, F., Yan, C., Zhou, D., & Zeng, X. (2022). An Efficient Batch-Constrained Bayesian Optimization Approach for Analog Circuit Synthesis via Multiobjective Acquisition Ensemble. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *41*(1), 1–14. https://doi.org/10.1109/TCAD.2021.3054811

[43] van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primers*, *1*(1), 1. https://doi.org/10.1038/s43586-020-00001-2

[44] Wang, X., Jin, Y., Schmitt, S., & Olhofer, M. (2023). Recent Advances in Bayesian Optimization. *ACM Computing Surveys*, *55*(13s), 1–36. https://doi.org/10.1145/3582078

[45] Agnihotri, A., & Batra, N. (2020). Exploring Bayesian Optimization. *Distill*, *5*(5). https://doi.org/10.23915/distill.00026

[46] Idroes, G. M., Maulana, A., Suhendra, R., Lala, A., Karma, T., Kusumo, F., Hewindati, Y. T., & Noviandy, T. R. (2023). TeutongNet: A Fine-Tuned Deep Learning Model for Improved Forest Fire Detection. *Leuser Journal of Environmental Studies*, *1*(1), 1–8. https://doi.org/10.60084/ljes.v1i1.42

[47] Noviandy, T. R., Alfanshury, M. H., Abidin, T. F., & Riza, H. (2023). Enhancing Glioma Grading Performance: A Comparative Study on Feature Selection Techniques and Ensemble Machine Learning. *2023 International Conference on Computer, Control, Informatics and Its Applications (IC3INA)*, 406–411. https://doi.org/10.1109/IC3INA60834.2023.10285778

[48] Maulana, A., Noviandy, T. R., Suhendra, R., Earlia, N., Bulqiah, M., Idroes, G. M., Niode, N. J., Sofyan, H., Subianto, M., & Idroes, R. (2023). Evaluation of Atopic Dermatitis Severity Using Artificial Intelligence. *Narra J*, *3*(3), e511. https://doi.org/10.52225/narra.v3i3.511