Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 1, No. 1, 2023

Ensemble Machine Learning Approach for Quantitative Structure-Activity Relationship Based Drug Discovery: A Review

Teuku Rizky Noviandy [1,2], Aga Maulana [1,2], Ghazi Mauer Idroes [3,4,*], Talha Bin Emran [5], Trina Ekawati Tallei [6], Zuchra Helwani [7] and Rinaldi Idroes [8]

[1] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; trizkynoviandy@gmail.com (T.R.N.); agamaulana@usk.ac.id (A.M.)
[2] Data Science and Artificial Intelligence Research Unit, Graha Primera Saintifika, Aceh Besar 23371, Indonesia;
[3] Graduate School of Mathematics and Applied Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; idroesghazi_k3@abulyatama.ac.id (G.M.I.)
[4] Department of Occupational Health and Safety, Faculty of Health Sciences, Universitas Abulyatama, Aceh Besar 23372, Indonesia;
[5] Department of Pharmacy, BGC Trust University Bangladesh, Chittagong 4381, Bangladesh; talhabmb@bgctub.ac.bd (T.B.E.)
[6] Department of Biology, Faculty of Mathematics and Natural Sciences, Sam Ratulangi University, Manado 95115, North Sulawesi, Indonesia; trina_tallei@unsrat.ac.id (T.E.T.)
[7] Department of Chemical Engineering, Universitas Riau Pekanbaru 28293, Indonesia; zuchra.helwani@lecturer.unri.ac.id (Z.H.)
[8] Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; rinaldi.idroes@usk.ac.id (R.I.)

***** Correspondence: idroesghazi_k3@abulyatama.ac.id

Article History

Received 10 August 2023
Revised 15 September 2023
Accepted 21 September 2023
Available Online 25 September 2023

Keywords:
QSAR
Ensemble techniques
Molecular descriptors
Classification
Regression

Abstract

This comprehensive review explores the pivotal role of ensemble machine learning techniques in Quantitative Structure-Activity Relationship (QSAR) modeling for drug discovery. It emphasizes the significance of accurate QSAR models in streamlining candidate compound selection and highlights how ensemble methods, including AdaBoost, Gradient Boosting, Random Forest, Extra Trees, XGBoost, LightGBM, and CatBoost, effectively address challenges such as overfitting and noisy data. The review presents recent applications of ensemble learning in both classification and regression tasks within QSAR, showcasing the exceptional predictive accuracy of these techniques across diverse datasets and target properties. It also discusses the key challenges and considerations in ensemble QSAR modeling, including data quality, model selection, computational resources, and overfitting. The review outlines future directions in ensemble QSAR modeling, including the integration of multi-modal data, explainability, handling imbalanced data, automation, and personalized medicine applications while emphasizing the need for ethical and regulatory guidelines in this evolving field.

1. Introduction

To discover new potential candidates during the drug design and discovery process, it is important to utilize efficient and reliable computational techniques [1–3]. One widely employed method is Quantitative Structure-Activity Relationship (QSAR) modeling [4]. In QSAR modeling, mathematical models are constructed to establish the relationship between the structural and chemical characteristics of molecules and their biological activities [5].

The importance of accurate QSAR models in drug discovery cannot be overstated. These models enable researchers to make informed decisions regarding the potential efficacy, safety, and toxicity of candidate compounds, thereby significantly reducing the time and resources required for experimental testing [6]. Accurate QSAR models also aid in the identification of lead compounds with desirable properties, paving the way for the development of innovative drugs and chemical solutions that can address pressing medical and industrial challenges [7].

As the pharmaceutical industry continues to grapple with escalating research costs and increasing pressure to deliver safe and effective drugs in a timely manner, the need for robust and accurate QSAR models has grown exponentially. For this purpose, machine learning approaches have gained prominence as a means to enhance the predictive performance and reliability of QSAR models [8, 9].

Machine learning is a branch of artificial intelligence (AI) that focuses on developing algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed [10, 11]. It involves using data to train these algorithms, allowing them to recognize patterns, make predictions, and improve their performance over time [12–14]. Machine learning's versatility and utility have become increasingly apparent as it discovers innovative and impactful applications in various fields [15–18].

One popular machine learning technique is ensemble machine learning. This technique combines multiple machine learning models to improve predictive accuracy and robustness. Instead of relying on a single model, ensemble techniques assemble a group of models that work together to make more accurate predictions, reduce the risk of overfitting, and enhance the overall performance of machine learning algorithms [19, 20]. Numerous studies conducted across various domains have consistently demonstrated the remarkable performance benefits of ensemble machine learning techniques [21–24].

This review aims to provide a comprehensive overview of ensemble machine learning techniques in QSAR-based drug discovery. We will delve into the fundamental concepts of QSAR modeling, discuss the challenges associated with traditional QSAR models, and elucidate how ensemble methods offer a promising avenue to address these challenges. Furthermore, we will explore the recent advancements and applications of ensemble techniques in drug discovery, highlighting their potential to accelerate the identification of novel drug candidates.

2. QSAR Modeling

2.1. Explanation of QSAR

QSAR is a fundamental concept in drug discovery and computational chemistry that involves predicting the biological activity of chemical compounds based on their structural features [25]. QSAR establishes a quantitative link between a molecule's structure and its biological activity through mathematical models to help understand how the molecular characteristics of a compound influence its effectiveness as a drug candidate [26].

In drug discovery, QSAR plays an important role in screening and prioritizing compounds for further development [27, 28]. It allows researchers to identify molecules with the desired biological activity and optimize their chemical structures to enhance potency and selectivity. Additionally, QSAR models aid in predicting the toxicity and safety profiles of compounds, which is crucial for minimizing adverse effects during clinical trials [29, 30]. Overall, QSAR has become an indispensable tool in modern drug discovery, facilitating the efficient and cost-effective identification of potential therapeutic agents.

QSAR tasks are typically categorized into two main types: classification and regression [31, 32]. Classification involves the prediction of discrete outcomes, such as whether a compound will exhibit a particular biological activity (e.g., active or inactive) [22]. QSAR classification models use machine learning algorithms to classify compounds into predefined categories based on their structural features and properties. These models are invaluable for early-stage drug discovery, as they help identify potential lead compounds and prioritize further experimental testing.

On the other hand, regression tasks in QSAR deal with predicting continuous numerical values, often associated with the degree or potency of a biological response. For instance, researchers may use regression models to estimate the IC_{50} (half-maximal inhibitory concentration) or EC_{50} (half-maximal effective concentration) values of compounds, which provide quantitative measures of their efficacy [33]. Regression-based QSAR models leverage mathematical algorithms to establish a relationship between the structural attributes of molecules and their corresponding quantitative biological activities [34]. These models are crucial for optimizing drug candidates and fine-tuning their chemical structures to achieve the desired level of potency.

2.2. Molecular Descriptors

Molecular descriptors are important components of QSAR studies that serve as the quantitative representations of a molecule's structural and chemical characteristics [35]. Molecular descriptors encompass a wide range of numerical values that capture various aspects of a molecule's structure and properties [36]. These descriptors can include information about molecular size, shape, electronic distribution, hydrophobicity, and many other attributes and are important for transforming complex molecular structures into numerical data that can be used in mathematical models [37].

Various specialized software tools and libraries have been developed to calculate molecular descriptors efficiently. These tools, such as Mordred [38], Dragon [39], RDKit [40], AlvaDesc [41], and PaDEL-descriptor [42], play an instrumental role in automating the process of descriptor calculation, saving researchers valuable time and ensuring accuracy.

Molecular descriptors are used to build quantitative models that relate these numerical representations to the biological activity or property of interest. By analyzing the relationship between descriptors and activity data, QSAR models can make predictions about the activity of new compounds, facilitating the prioritization of potential drug candidates for further experimental evaluation. Consequently, the accurate selection and calculation of molecular descriptors are important in the successful application of QSAR techniques in drug discovery.

The integration of feature selection methods such as Genetic Algorithm (GA) [43, 44], Particle Swarm Optimization (PSO) [45], and Recursive Feature Elimination (RFE) [46] further enhances the effectiveness of QSAR models, allowing the identification of the most relevant molecular descriptors, allowing for a more focused and precise modeling process. This ensures that the QSAR model is built on the most informative features, ultimately improving its predictive accuracy and interpretability.

3. Ensemble Learning Techniques for QSAR

Machine learning for QSAR tasks often encounters challenges such as overfitting, model instability, and the presence of noisy or incomplete data [47, 48]. These issues can result in unreliable predictions and reduced model generalization. Ensemble learning, a powerful technique in the machine learning toolbox, offers a promising solution to mitigate these problems. By combining multiple base models, ensemble methods can enhance predictive accuracy, reduce variance, and improve the overall robustness of QSAR models. In this section, we delve into various ensemble learning techniques.

3.1. AdaBoost

AdaBoost, short for Adaptive Boosting, is a powerful ensemble method that combines multiple weak learners to create a strong, accurate model. It works by assigning weights to each data point and adjusting these weights during each iteration to focus on the instances that were previously misclassified. Weak learners, often decision trees with limited depth, are trained sequentially, and their predictions are combined to form the final ensemble model. AdaBoost is particularly effective when dealing with complex classification problems and can adapt to various data distributions, making it a valuable tool in machine learning for improving predictive performance [49, 50].

3.2. Gradient Boosting

Gradient Boosting is a versatile ensemble method that builds a strong predictive model by sequentially training a series of decision trees [51]. Unlike AdaBoost, Gradient Boosting focuses on minimizing the errors made by the previous trees at each step. It does this by fitting new trees to the residuals or errors of the previous predictions. This iterative process helps Gradient Boosting gradually refine its model, making it adept at handling both regression and classification tasks [52].

3.3. Random Forest

Random Forest is a robust and versatile algorithm that leverages the power of decision trees to create an ensemble model [53]. However, it differs from traditional decision tree models by constructing multiple trees during training. These trees are built using bootstrapped samples of the data, and at each node, a random subset of features is considered for splitting. This randomness helps reduce overfitting and makes Random Forest less susceptible to outliers [54]. The final prediction is obtained by aggregating the predictions of all individual trees, typically through majority voting for classification or averaging for regression. Random Forest is highly effective, easy to implement, and resistant to overfitting, making it a popular choice for a wide range of machine learning tasks [46, 55].

3.4. Extra Trees

Extra Trees, also known as Extremely Randomized Trees, is an advanced ensemble learning technique that extends the principles of Random Forest. While Random Forest selects the best split among a random subset of features at each node, Extra Trees goes a step further by making

completely random splits. This additional randomness makes Extra Trees even less prone to overfitting than Random Forest [56]. By building multiple randomized trees and combining their predictions through averaging (for regression) or majority voting (for classification), Extra Trees offers a powerful and highly robust approach and is useful when dealing with noisy or high-dimensional datasets [57].

3.5. XGBoost

XGBoost (Extreme Gradient Boosting) is an advanced gradient boosting algorithm that works by improving traditional gradient boosting methods with regularization techniques, parallel processing, and a clever optimization algorithm [58]. XGBoost can handle both regression and classification tasks and has the ability to work with missing data effectively. It is highly customizable, allowing users to fine-tune various hyperparameters to achieve optimal results for their specific problem. XGBoost has gained widespread popularity and utility in both machine learning competitions and real-world applications due to its exceptional versatility and high-performance capabilities [23, 59].

3.6. LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework developed by Microsoft. It distinguishes itself through its speed and efficiency, making it particularly well-suited for large datasets and high-dimensional feature spaces [60]. LightGBM employs a histogram-based approach to find the best splits during tree construction, which drastically reduces computation time [61]. With its impressive scalability, customizable options, and ability to handle categorical features efficiently, LightGBM has become a preferred choice for many data scientists and machine learning practitioners seeking top-tier performance in diverse applications [21, 62, 63].

3.7. CatBoost

CatBoost is a machine learning algorithm that excels in dealing with datasets containing categorical variables [64]. Unlike traditional gradient boosting methods that require one-hot encoding or other preprocessing techniques for categorical features, CatBoost can work directly with such data [65]. It employs a technique called ordered boosting, which optimizes the order of categorical features during tree construction, reducing overfitting and improving predictive accuracy. Additionally, CatBoost incorporates robust handling of missing values and offers efficient GPU support for accelerated training. Its ease of use and exceptional performance on a wide range of tasks have made CatBoost a valuable tool in the machine learning toolkit, especially for those dealing with real-world datasets rich in categorical information [24, 66].

4. Applications of Ensemble Learning in QSAR

In this study, we conducted a literature review spanning the last five years, with a specific focus on the application of ensemble learning techniques in QSAR for drug discovery.

4.1. Classification Task

Table 1. provides a comprehensive overview of recent applications of ensemble machine learning techniques in QSAR for classification tasks. Ensemble methods, such as Random Forest, XGBoost, AdaBoost, and Gradient Boosting, have consistently demonstrated their prowess in improving predictive accuracy when compared to individual machine learning models. These ensemble approaches have been applied to tackle a diverse range of challenges, from predicting drug interactions with specific protein targets like Beta secretase 1 or Bruton's tyrosine kinase to understanding the behavior of viral proteins in diseases like Plasmodium falciparum or SARS-Cov-2.

Notably, XGBoost stands out as a recurrently successful ensemble method, consistently achieving high accuracy rates across multiple studies. In the context of Bruton's tyrosine kinase inhibitor prediction, for instance, XGBoost achieved an exceptional accuracy rate of 94.1%, surpassing all other machine learning models. In addition to XGBoost, Random Forest has also proven to be a formidable choice, particularly in applications related to drug discovery. These ensemble techniques harness the collective intelligence of multiple base models, allowing them to capture complex relationships within the data and enhance predictive performance.

Furthermore, the table underscores the importance of selecting the most suitable ensemble method for each specific problem. Different datasets and target proteins may benefit from distinct ensemble strategies, as evidenced by the variations in model performance across studies. As machine learning continues to play a pivotal role in advancing drug discovery and molecular biology, the effective utilization of ensemble methods remains a promising avenue for improving the accuracy and reliability of predictive models in these critical domains.

4.2. Regression Task

Table 2. presents the application of ensemble methods for regression tasks in QSAR for drug discovery. In these studies, Random Forest consistently emerges as a powerful ensemble method, demonstrating its versatility

Table 1. Application of ensemble machine learning model in QSAR for classification task.

Method	Year	Dataset	Findings	Ref.
SVM, KNN, Random Forest, XGBoost	2019	*Plasmodium falciparum*	the XGBoost model achieved an accuracy rate of 86.00%, outperforming all machine learning models.	[67]
Random Forest, AdaBoost, Extra Trees	2020	Dengue virus NS3 protein	The Extra Trees model achieved the highest performance with an accuracy score of 73.00%, outperforming other machine learning models.	[68]
Naïve Bayesian, KNN, SVM, Random Forest, XGBoost	2021	Beta secretase 1	Random Forest and XGBoost achieved F1-score of 87.00%, respectively	[69]
Random Forest, XGBoost, Naïve Bayesian, SVM, ANN	2022	SARS-Cov-2	The XGBoost model achieved an accuracy rate of 84.50%, exceeding the accuracy levels of all other machine learning models.	[70]
Random Forest, XGBoost, Decision Tree, QDA, KNN, SVM, LDA, LR, Naïve Bayesian, ANN	2022	3CLPro-protease inhibitor	XGBoost and Random Forest achieved the highest accuracy compared to the other machine learning models.	[71]
AdaBoost, Extra Trees, Gradient Boosting, Random Forest,	2023	Beta secretase 1 inhibitor	The Random Forest model attained an accuracy of 82.53%, surpassing the accuracy of all other machine learning models.	[72]
Random Forest, Gradient Boosting, LightGBM, XGBoost, Extra Trees	2023	Androgen receptor	The Extra Trees model achieved an accuracy of 73.50%, which was the highest among all the other machine learning models.	[73]
LightGBM	2023	Acetylcholinesterase inhibitor	The LightGBM model achieved an accuracy of 82.47%	[62]
AdaBoost, Gradient Boosting, Random Forest	2023	Diacylglycerol Acyltransferase-1 inhibitor	The gradient boosting model attained the highest accuracy of 80.00% compared to the other machine learning models.	[74]
Random Forest, XGBoost	2023	Bruton's tyrosine kinase inhibitor	The XGBoost model attained an accuracy of 94.1%, surpassing all other machine learning models in terms of performance.	[75]
LightGBM, XGBoost	2023	Hepatitis C NS5B protein	The combination of LightGBM and XGBoost achieved the highest accuracy of accuracy of 85.07% compared to the individual model.	[76]

Table 2. Application of ensemble machine learning model in QSAR for regression task.

Ensemble Method	Year	Dataset	Findings	Ref.
Decision Tree, SVM, DNN, Random Forest	2019	Janus kinase 2	Random Forest achieved the highest R^2 of 0.74 compared to the other machine learning models.	[77]
Linear Regression, Random Forest, SVM	2019	5-lipoxygenase inhibitors	Random Forest achieved the highest R^2 of 0.93 compared to the other machine learning models.	[78]
SVM, Random Forest, ANN, KNN, Deep Learning, Linear Regression	2020	p38α mitogen-activated protein kinase inhibitors	Random Forest achieved the highest R^2 of 0.82 compared to the other machine learning models.	[79]
XGBoost	2020	Dipeptidyl peptidase-4	XGBoost achieved the highest R^2 of 0.94	[80]
Random Forest	2021	STAT3	Random Forest achieved an R^2 of 0.886	[81]
Random Forest	2022	Hepatitis C NS5B inhibitors	Random Forest achieved the highest R^2 of 0.73	[33]
Catboost, LightGBM, XGBoost	2023	p-glycoprotein inhibitors	Catboost, LightGBM, and XGBoost achieved R^2 of 0.97, respectively	[82]
XGBoost	2023	Kirsten rat sarcoma viral G12C	XGBoost achieved an R^2 of 0.76	[83]

and robustness across different datasets and target variables. For instance, in the context of Janus kinase 2 prediction, Random Forest achieved an impressive R^2 value of 0.74, while in the study of 5-lipoxygenase inhibitors, it achieved an exceptionally high R^2 of 0.93. Random Forest also performed well in predicting the activity of STAT3, Hepatitis C NS5B inhibitors, and other targets, further highlighting its reliability in regression tasks.

XGBoost, another ensemble technique, also showcased its prowess in several instances. For example, in the prediction of Dipeptidyl peptidase-4 activity, XGBoost achieved an outstanding R^2 of 0.94. Similarly, in the case of Kirsten rat sarcoma viral G12C prediction, XGBoost achieved a commendable R^2 of 0.76.

Moreover, the table introduces the effectiveness of newer ensemble methods, such as CatBoost and

LightGBM, which demonstrated their promise in modeling p-glycoprotein inhibitors with an R^2 of 0.97 each. This suggests that the ensemble landscape for regression tasks is continually evolving, with emerging methods offering competitive performance.

These findings collectively underscore the significance of ensemble methods in regression-based drug discovery and molecular biology research. The selection of the most suitable ensemble method often depends on the specific dataset and target property under investigation, emphasizing the importance of tailoring the modeling approach to the unique characteristics of each problem. As regression tasks continue to play a pivotal role in understanding molecular interactions and designing novel drugs, the application of ensemble techniques remains a valuable strategy for enhancing predictive accuracy and advancing scientific discovery.

5. Challenges and Considerations in Ensemble QSAR Modeling

Ensemble QSAR modeling, while a powerful approach for enhancing predictive accuracy, has its challenges and considerations. In this section, we explore some of the key issues that researchers and practitioners must address when employing ensemble techniques in QSAR modeling.

5.1. Data Quality and Preprocessing

Ensemble models are sensitive to the quality and preprocessing of input data. Challenges arise when dealing with noisy, incomplete, or biased datasets. Data preprocessing, including feature selection, handling missing values, and addressing class imbalance, becomes critical to ensure the effectiveness of ensemble techniques [84, 85].

5.2. Model Selection and Hyperparameter Tuning

Choosing the right ensemble algorithm and configuring its hyperparameters can be a complex task. Different ensemble methods may perform better on specific datasets or for particular QSAR problems. Additionally, tuning the hyperparameters for each base model within the ensemble requires careful consideration to achieve optimal results [86].

5.3. Computational Resources

Many ensemble techniques, especially gradient boosting methods like XGBoost and LightGBM, can be computationally intensive [87]. Training multiple models in parallel or sequentially may require substantial computational resources and time, making it essential to balance model complexity with available resources.

5.4. Overfitting

Ensemble models, if not appropriately regularized or if the base models are too complex, can be susceptible to overfitting. Overfit models perform well on training data but poorly on unseen data, undermining the generalization ability of the ensemble [88]. Regularization techniques and cross-validation are crucial for mitigating overfitting risks [89, 90].

6. Future Directions in Ensemble QSAR Modeling

Ensemble QSAR modeling, a robust approach for improving predictive accuracy in drug discovery and molecular modeling, is poised for significant advancements in the future. One promising direction is the integration of multi-modal data, encompassing molecular structures, omics data, and clinical information, to provide a more comprehensive understanding of drug-target interactions and enhance prediction accuracy [91]. Additionally, the fusion of traditional ensemble methods with deep learning approaches offers exciting prospects, combining the representation learning capabilities of deep models with the reliability of ensembles [92].

In order to get more interpretable models, the development of explainable ensemble techniques will remain a priority. These models will elucidate the contributions of individual models within the ensemble, addressing regulatory compliance and facilitating decision-making in drug development [93, 94]. Leveraging transfer learning and pre-trained models will reduce the computational cost of training ensembles and accelerate model development, particularly for specialized QSAR tasks.

Furthermore, ensemble QSAR models are expected to play a pivotal role in real-time drug discovery and personalized medicine [95]. Tailoring drug recommendations and treatment plans to individual patients' genetic and molecular profiles will rely on highly accurate and adaptable ensemble models [96]. As the field advances, ethical and regulatory considerations will also become increasingly significant, necessitating the establishment of guidelines to ensure fairness, transparency, and accountability in model development and deployment.

7. Conclusion

In summary, this review paper highlights the significant role of ensemble machine learning techniques in advancing QSAR modeling for drug discovery. These methods offer enhanced predictive accuracy and robustness, addressing challenges in traditional QSAR

models. Recent applications demonstrate their success in providing higher performance compared to classic machine learning models. However, challenges such as data quality, model selection, and overfitting must be carefully managed. The future promises the integration of multi-modal data, explainable models, and personalized medicine applications. Ensemble QSAR modeling remains a vital tool in accelerating drug discovery and addressing medical challenges.

Author Contributions: Conceptualization, T.R.N., A.M., G.M.I., and R.I.; writing—original draft preparation, T.R.N., A.M., T.B.E., T.E.T.; writing—review and editing, G.M.I., Z.H., and R.I.; supervision, G.M.I., and R.I.; project administration, G.M.I. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Conflicts of Interest: All the authors declare that there are no conflicts of interest.

References

1. Golbraikh, A., Wang, X. S., Zhu, H., and Tropsha, A. (2017). Predictive QSAR Modeling: Methods and Applications in Drug Discovery and Chemical Risk Assessment, *Handbook of Computational Chemistry*, Springer International Publishing, Cham, 2303–2340. doi:10.1007/978-3-319-27282-5_37.
2. Maulydia, N. B., Khairan, K., and 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*, Vol. 1, No. 1, 16–21. doi:10.60084/mp.v1i1.33.
3. Khairan, K., Idroes, R., Tumilaar, S. G., Tallei, T. E., Idroes, G. M., Rahmadhany, F., Futri, M. U., Dinura, N. M., Mauliza, S., Diana, M., Maisarah, C. P., Maulana, A., Noviandy, T. R., Suhendra, R., Muslem, and Earlia, N. (2021). Molecular docking study of fatty acids from Pliek U Oil in the inhibition of SARS-CoV-2 protein and enzymes, *IOP Conference Series: Materials Science and Engineering*, Vol. 1087, No. 1, 012058. doi:10.1088/1757-899X/1087/1/012058.
4. Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V, Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., Isayev, O., Curtalolo, S., Fourches, D., Cohen, Y., Aspuru-Guzik, A., Winkler, D. A., Agrafiotis, D., Cherkasov, A., and Tropsha, A. (2020). QSAR without borders, *Chemical Society Reviews*, Vol. 49, No. 11, 3525–3564. doi:10.1039/D0CS00098A.
5. Toropov, A. A., and Toropova, A. P. (2020). QSPR/QSAR: State-of-Art, Weirdness, the Future, *Molecules*, Vol. 25, No. 6, 1292. doi:10.3390/molecules25061292.
6. Shen, J., and Nicolaou, C. A. (2019). Molecular property prediction: recent trends in the era of artificial intelligence, *Drug Discovery Today: Technologies*, Vols 32–33, 29–36. doi:10.1016/j.ddtec.2020.05.001.
7. Sabe, V. T., Ntombela, T., Jhamba, L. A., Maguire, G. E. M., Govender, T., Naicker, T., and Kruger, H. G. (2021). Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review, *European Journal of Medicinal Chemistry*, Vol. 224, 113705. doi:10.1016/j.ejmech.2021.113705.
8. Kwon, S., Bae, H., Jo, J., and Yoon, S. (2019). Comprehensive ensemble in QSAR prediction for drug discovery, *BMC Bioinformatics*, Vol. 20, No. 1, 521. doi:10.1186/s12859-019-3135-4.
9. Staszak, M., Staszak, K., Wieszczycka, K., Bajek, A., Roszkowski, K., and Tylkowski, B. (2022). Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship, *WIREs Computational Molecular Science*, Vol. 12, No. 2. doi:10.1002/wcms.1568.
10. Mahesh, B. (2020). Machine learning algorithms-a review, *International Journal of Science and Research (IJSR).[Internet]*, Vol. 9, No. 1, 381–386.
11. Hamet, P., and Tremblay, J. (2017). Artificial intelligence in medicine, *Metabolism*, Vol. 69, S36–S40. doi:10.1016/j.metabol.2017.01.011.
12. Kang, J., Schwartz, R., Flickinger, J., and Beriwal, S. (2015). Machine Learning Approaches for Predicting Radiation Therapy Outcomes: A Clinician's Perspective, *International Journal of Radiation Oncology*Biology*Physics*, Vol. 93, No. 5, 1127–1135. doi:10.1016/j.ijrobp.2015.07.2286.
13. Noviandy, T. R., Maulana, A., Idroes, G. M., Suhendra, R., Adam, M., Rusyana, A., and Sofyan, H. (2023). Deep Learning-Based Bitcoin Price Forecasting Using Neural Prophet, *Ekonomikalia Journal of Economics*, Vol. 1, No. 1, 19–25. doi:10.60084/eje.v1i1.51.
14. Idroes, G. M., Maulana, A., Suhendra, R., Lala, A., Karma, T., Kusumo, F., Hewindati, Y. T., and Noviandy, T. R. (2023). TeutongNet: A Fine-Tuned Deep Learning Model for Improved Forest Fire Detection, *Leuser Journal of Environmental Studies*, Vol. 1, No. 1, 1–8. doi:10.60084/ljes.v1i1.42.
15. Maulana, A., Faisal, F. R., Noviandy, T. R., Rizkia, T., Idroes, G. M., Tallei, T. E., El-Shazly, M., and Idroes, R. (2023). Machine Learning Approach for Diabetes Detection Using Fine-Tuned XGBoost Algorithm, *Infolitika Journal of Data Science*, Vol. 1, No. 1, 1–7. doi:10.60084/ijds.v1i1.72.
16. 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., and 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)*, IEEE, 13–18. doi:10.1109/ICELTICs56128.2022.9932124.
17. Idroes, R., Noviandy, T. R., Maulana, A., Suhendra, R., Sasmita, N. R., Muslem, M., Idroes, G. M., Kemala, P., and 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)*, Vol. 14, No. 2, 137. doi:10.15866/iremos.v14i2.20460.
18. Maulana, A., Noviandy, T. R., Sasmita, N. R., Paristiowati, M., Suhendra, R., Yandri, E., Satrio, J., and Idroes, R. (2023). Optimizing University Admissions: A Machine Learning Perspective, *Journal of Educational Management and Learning*, Vol. 1, No. 1, 1–7. doi:10.60084/jeml.v1i1.46.
19. Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning, *Frontiers of Computer Science*, Vol. 14, No. 2, 241–258. doi:10.1007/s11704-019-8208-z.
20. Sagi, O., and Rokach, L. (2018). Ensemble learning: A survey, *WIREs Data Mining and Knowledge Discovery*, Vol. 8, No. 4. doi:10.1002/widm.1249.
21. Rufo, D. D., Debelee, T. G., Ibenthal, A., and Negera, W. G. (2021). Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM), *Diagnostics*, Vol. 11, No. 9, 1714. doi:10.3390/diagnostics11091714.
22. Simeon, S., Anuwongcharoen, N., Shoombuatong, W., Malik, A. A., Prachayasittikul, V., Wikberg, J. E. S., and Nantasenamat, C. (2016). Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking, *PeerJ*, Vol. 4, e2322. doi:10.7717/peerj.2322.
23. Amjad, M., Ahmad, I., Ahmad, M., Wróblewski, P., Kamiński, P., and Amjad, U. (2022). Prediction of pile bearing capacity using XGBoost algorithm: modeling and performance evaluation, *Applied Sciences*, Vol. 12, No. 4, 2126.

24. Kumar, P. S., K, A. K., Mohapatra, S., Naik, B., Nayak, J., and Mishra, M. (2021). CatBoost Ensemble Approach for Diabetes Risk Prediction at Early Stages, *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology(ODICON)*, IEEE, 1–6. doi:10.1109/ODICON50556.2021.9428943.

25. Tropsha, A. (2010). Best Practices for QSAR Model Development, Validation, and Exploitation, *Molecular Informatics*, Vol. 29, Nos. 6–7, 476–488. doi:10.1002/minf.201000061.

26. Puzyn, T., Leszczyński, J., and Cronin, M. (2010). *Recent Advances in QSAR Studies*, (T. Puzyn, J. Leszczynski, & M. T. Cronin, Eds.)... *Advances in Computational Chemistry …* (Vol. 8), Springer Netherlands, Dordrecht. doi:10.1007/978-1-4020-9783-6.

27. Tropsha, A., and Golbraikh, A. (2007). Predictive QSAR Modeling Workflow, Model Applicability Domains, and Virtual Screening, *Current Pharmaceutical Design*, Vol. 13, No. 34, 3494–3504. doi:10.2174/138161207782794257.

28. Neves, B. J., Braga, R. C., Melo-Filho, C. C., Moreira-Filho, J. T., Muratov, E. N., and Andrade, C. H. (2018). QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery, *Frontiers in Pharmacology*, Vol. 9. doi:10.3389/fphar.2018.01275.

29. Abramenko, N., Kustov, L., Metelytsia, L., Kovalishyn, V., Tetko, I., and Peijnenburg, W. (2020). A review of recent advances towards the development of QSAR models for toxicity assessment of ionic liquids, *Journal of Hazardous Materials*, Vol. 384, 121429. doi:10.1016/j.jhazmat.2019.121429.

30. Fan, F., Toledo Warshaviak, D., Hamadeh, H. K., and Dunn, R. T. (2019). The integration of pharmacophore-based 3D QSAR modeling and virtual screening in safety profiling: A case study to identify antagonistic activities against adenosine receptor, A2A, using 1,897 known drugs, *PLOS ONE*, Vol. 14, No. 1, e0204378. doi:10.1371/journal.pone.0204378.

31. Pirhadi, S., Shiri, F., and Ghasemi, J. B. (2015). Multivariate statistical analysis methods in QSAR, *Rsc Advances*, Vol. 5, No. 127, 104635–104665.

32. Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T. D., McDowell, R. M., and Gramatica, P. (2003). Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs., *Environmental Health Perspectives*, Vol. 111, No. 10, 1361–1375. doi:10.1289/ehp.5758.

33. Kamboj, S., Rajput, A., Rastogi, A., Thakur, A., and Kumar, M. (2022). Targeting non-structural proteins of Hepatitis C virus for predicting repurposed drugs using QSAR and machine learning approaches, *Computational and Structural Biotechnology Journal*, Vol. 20, 3422–3438. doi:10.1016/j.csbj.2022.06.060.

34. Yang, B., Si, H., and Zhai, H. (2021). QSAR Studies on the IC50 of a Class of Thiazolidinone/Thiazolide Based Hybrids as Antitrypanosomal Agents, *Letters in Drug Design & Discovery*, Vol. 18, No. 4, 406–415. doi:10.2174/1570180817999201102200015.

35. Todeschini, R., and Consonni, V. (2000). *Handbook of Molecular Descriptors*, Wiley-VCH Verlag GmbH, Weinheim, Germany. doi:10.1002/9783527613106.

36. Mauri, A., Consonni, V., and Todeschini, R. (2017). Molecular Descriptors, *Handbook of Computational Chemistry*, Springer International Publishing, Cham, 2065–2093. doi:10.1007/978-3-319-27282-5_51.

37. Xue, L., and Bajorath, J. (2000). Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening, *Combinatorial Chemistry & High Throughput Screening*, Vol. 3, No. 5, 363–372. doi:10.2174/1386207003331454.

38. Moriwaki, H., Tian, Y. S., Kawashita, N., and Takagi, T. (2018). Mordred: A molecular descriptor calculator, *Journal of Cheminformatics*, Vol. 10, No. 1, 1–14. doi:10.1186/s13321-018-0258-y.

39. Mauri, A., Consonni, V., Pavan, M., and Todeschini, R. (2006). Dragon software: An easy approach to molecular descriptor calculations, *Match*, Vol. 56, No. 2, 237–248.

40. Landrum, G. (2016). Rdkit: Open-source cheminformatics software.

41. Mauri, A. (2020). alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints, 801–820. doi:10.1007/978-1-0716-0150-1_32.

42. Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints, *Journal of Computational Chemistry*, Vol. 32, No. 7, 1466–1474. doi:10.1002/jcc.21707.

43. Noviandy, T. R., Maulana, A., Sasmita, N. R., Suhendra, R., Irvanizam, I., Muslem, M., Idroes, G. M., Yusuf, M., Sofyan, H., Abidin, T. F., and 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*, Vol. 17, No. 1, 306–326.

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

45. Ramaswamy, R., Kandhasamy, P., and Palaniswamy, S. (2023). Feature Selection for Alzheimer's Gene Expression Data Using Modified Binary Particle Swarm Optimization, *IETE Journal of Research*, Vol. 69, No. 1, 9–20. doi:10.1080/03772063.2021.1962747.

46. Bahl, A., Hellack, B., Balas, M., Dinischiotu, A., Wiemann, M., Brinkmann, J., Luch, A., Renard, B. Y., and Haase, A. (2019). Recursive feature elimination in random forest classification supports nanomaterial grouping, *NanoImpact*, Vol. 15, 100179. doi:10.1016/j.impact.2019.100179.

47. Ying, X. (2019). An Overview of Overfitting and its Solutions, *Journal of Physics: Conference Series*, Vol. 1168, 022022. doi:10.1088/1742-6596/1168/2/022022.

48. Yang, X., Wang, Y., Byrne, R., Schneider, G., and Yang, S. (2019). Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery, *Chemical Reviews*, Vol. 119, No. 18, 10520–10594. doi:10.1021/acs.chemrev.8b00728.

49. Ying, C., Qi-Guang, M., Jia-Chen, L., and Lin, G. (2013). Advance and prospects of AdaBoost algorithm, *Acta Automatica Sinica*, Vol. 39, No. 6, 745–758.

50. Cao, Y., Miao, Q.-G., Liu, J.-C., and Gao, L. (2013). Advance and Prospects of AdaBoost Algorithm, *Acta Automatica Sinica*, Vol. 39, No. 6, 745–758. doi:10.1016/S1874-1029(13)60052-X.

51. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine., *The Annals of Statistics*, Vol. 29, No. 5. doi:10.1214/aos/1013203451.

52. Natekin, A., and Knoll, A. (2013). Gradient boosting machines, a tutorial, *Frontiers in Neurorobotics*, Vol. 7, 21. doi:10.3389/fnbot.2013.00021.

53. Biau, G., and Scornet, E. (2016). A random forest guided tour, *TEST*, Vol. 25, No. 2, 197–227. doi:10.1007/s11749-016-0481-7.

54. Qi, Y. (2012). Random forest for bioinformatics, *Ensemble Machine Learning: Methods and Applications*, Springer, 307–323.

55. Edeh, M. O., Khalaf, O. I., Tavera, C. A., Tayeb, S., Ghouali, S., Abdulsahib, G. M., Richard-Nnabu, N. E., and Louni, A. (2022). A Classification Algorithm-Based Hybrid Diabetes Prediction Model, *Frontiers in Public Health*, Vol. 10. doi:10.3389/fpubh.2022.829519.

56. Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees, *Machine Learning*, Vol. 63, No. 1, 3–42. doi:10.1007/s10994-006-6226-1.

57. Goetz, M., Weber, C., Bloecher, J., Stieltjes, B., Meinzer, H.-P., and Maier-Hein, K. (2014). Extremely randomized trees based brain tumor segmentation, *Proceeding of BRATS Challenge-MICCAI*, Vol. 14, 6–11.

58. Chen, T., and 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.

59. Li, M., Fu, X., and Li, D. (2020). Diabetes Prediction Based on XGBoost Algorithm, *IOP Conference Series: Materials Science and Engineering*, Vol. 768, No. 7, 072093. doi:10.1088/1757-899X/768/7/072093.

60. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, Vol. 30.

61. Chen, C., Zhang, Q., Ma, Q., and Yu, B. (2019). LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion, *Chemometrics and Intelligent Laboratory Systems*, Vol. 191, 54–64. doi:10.1016/j.chemolab.2019.06.003.

62. Noviandy, T. R., Maulana, A., Idroes, G. M., Maulydia, N. B., Patwekar, M., Suhendra, R., and Idroes, R. (2023). Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer's Disease Drug Discovery, *Malacca Pharmaceutics*, Vol. 1, No. 2, 48–54. doi:10.60084/mp.v1i2.60.

63. Yang, H., Chen, Z., Yang, H., and Tian, M. (2023). Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison, *IEEE Access*, Vol. 11, 23366–23380. doi:10.1109/ACCESS.2023.3253885.

64. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features, *Advances in Neural Information Processing Systems*, Vol. 31.

65. Dorogush, A. V., Ershov, V., and Gulin, A. (2018). CatBoost: gradient boosting with categorical features support, *ArXiv Preprint ArXiv:1810.11363*.

66. Dhananjay, B., and Sivaraman, J. (2021). Analysis and classification of heart rate using CatBoost feature ranking model, *Biomedical Signal Processing and Control*, Vol. 68, 102610. doi:10.1016/j.bspc.2021.102610.

67. Danishuddin, Madhukar, G., Malik, M. Z., and Subbarao, N. (2019). Development and rigorous validation of antimalarial predictive models using machine learning approaches, *SAR and QSAR in Environmental Research*, Vol. 30, No. 8, 543–560. doi:10.1080/1062936X.2019.1635526.

68. Kurniawan, I., Rosalinda, M., and Ikhsan, N. (2020). Implementation of ensemble methods on QSAR Study of NS3 inhibitor activity as anti-dengue agent, *SAR and QSAR in Environmental Research*, Vol. 31, No. 6, 477–492.

69. Singh, R., Ganeshpurkar, A., Ghosh, P., Pokle, A. V., Kumar, D., Singh, R. bhushan, Singh, S. K., and Kumar, A. (2021). Classification of beta-site amyloid precursor protein cleaving enzyme 1 inhibitors by using machine learning methods, *Chemical Biology & Drug Design*, Vol. 98, No. 6, 1079–1097. doi:10.1111/cbdd.13965.

70. Azizah, M., Yanuar, A., and Firdayani, F. (2022). Dimensional Reduction of QSAR Features Using a Machine Learning Approach on the SARS-Cov-2 Inhibitor Database, *Jurnal Penelitian Pendidikan IPA*, Vol. 8, No. 6, 3095–3101. doi:10.29303/jppipa.v8i6.2432.

71. Mondal, K., and S, S. K. (2021). QSAR Classification Models for Predicting 3CLPro-protease Inhibitor Activity, *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*, IEEE, 1–6. doi:10.1109/GUCON50781.2021.9573896.

72. Noviandy, T. R., Maulana, A., Emran, T. B., Idroes, G. M., and 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*, Vol. 1, No. 1, 1–7. doi:10.60084/hjas.v1i1.12.

73. Yu, T., Nantasenamat, C., Kachenton, S., Anuwongcharoen, N., and Piacham, T. (2023). Cheminformatic Analysis and Machine Learning Modeling to Investigate Androgen Receptor Antagonists to Combat Prostate Cancer, *ACS Omega*, Vol. 8, No. 7, 6729–6742. doi:10.1021/acsomega.2c07346.

74. Arifa, I., Aditsania, A., and Kurniawan, I. (2023). The Implementation of Genetic Algorithm-Ensemble Learning on QSAR Study of Diacylglycerol Acyltransferase-1(DGAT1) Inhibitors as Anti-diabetes, 282–292. doi:10.1007/978-981-99-0741-0_20.

75. Li, G., Li, J., Tian, Y., Zhao, Y., Pang, X., and Yan, A. (2023). Machine learning-based classification models for non-covalent Bruton's tyrosine kinase inhibitors: predictive ability and interpretability, *Molecular Diversity*. doi:10.1007/s11030-023-10696-6.

76. Noviandy, T. R., Maulana, A., Idroes, G. M., Irvanizam, I., Subianto, M., and 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)*, IEEE, 220–225. doi:10.1109/COSITE60233.2023.10250039.

77. Simeon, S., and Jongkon, N. (2019). Construction of Quantitative Structure Activity Relationship (QSAR) Models to Predict Potency of Structurally Diversed Janus Kinase 2 Inhibitors, *Molecules*, Vol. 24, No. 23, 4393. doi:10.3390/molecules24234393.

78. Shameera Ahamed, T. K., Rajan, V. K., and Muraleedharan, K. (2019). QSAR modeling of benzoquinone derivatives as 5-lipoxygenase inhibitors, *Food Science and Human Wellness*, Vol. 8, No. 1, 53–62. doi:10.1016/j.fshw.2019.02.001.

79. Joel, I. Y., Adigun, T. O., Bankole, O. O., Iduze, M. A., AbelJack-Soala, T., ANI, O. G., Olapade, E. O., Dada, F. M., Adetiwa, O. M., Ofeniforo, B. E., and Akanni, F. O. (2020). Insights into features and lead optimization of novel type 1½ inhibitors of p38α mitogen-activated protein kinase using QSAR, quantum mechanics, bioisostere replacement and ADMET studies, *Results in Chemistry*, Vol. 2, 100044. doi:10.1016/j.rechem.2020.100044.

80. Husna, N. A., Bustamam, A., Yanuar, A., Sarwinda, D., and Hermansyah, O. (2020). The comparison of machine learning methods for prediction study of type 2 diabetes mellitus's drug design, 030010. doi:10.1063/5.0024161.

81. Patterson, J. M., Milligan, K., and Winstead, C. (2021). Development of a QSAR model to predict molecular inhibition of human STAT3, *BioRxiv*, 2010–2021. doi:10.1101/2021.10.29.466511.

82. Lahyaoui, M., Diane, A., El-Idrissi, H., Saffaj, T., Rodi, Y. K., and Ihssane, B. (2023). QSAR modeling and molecular docking studies of 2-oxo-1, 2-dihydroquinoline-4- carboxylic acid derivatives as p-glycoprotein inhibitors for combating cancer multidrug resistance, *Heliyon*, Vol. 9, No. 1, e13020. doi:10.1016/j.heliyon.2023.e13020.

83. Srisongkram, T., and Weerapreeyakul, N. (2022). Drug Repurposing against KRAS Mutant G12C: A Machine Learning, Molecular Docking, and Molecular Dynamics Study, *International Journal of Molecular Sciences*, Vol. 24, No. 1, 669. doi:10.3390/ijms24010669.

84. Noviandy, T. R., Idroes, G. M., Maulana, A., Hardi, I., Ringga, E. S., and Idroes, R. (2023). Credit Card Fraud Detection for Contemporary Financial Management Using XGBoost-Driven Machine Learning and Data Augmentation Techniques, *Indatu Journal of Management and Accounting*, Vol. 1, No. 1, 29–35.

85. Li, J., Luo, D., Wen, T., Liu, Q., and Mo, Z. (2021). Representative feature selection of molecular descriptors in QSAR modeling, *Journal of Molecular Structure*, Vol. 1244, 131249. doi:10.1016/j.molstruc.2021.131249.

86. Kurniawan, I., Rosalinda, M., and Ikhsan, N. (2020). Implementation of ensemble methods on QSAR Study of NS3 inhibitor activity as anti-dengue agent, *SAR and QSAR in Environmental Research*, Vol. 31, No. 6, 477–492. doi:10.1080/1062936X.2020.1773534.

87. Bentéjac, C., Csörgő, A., and Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms, *Artificial Intelligence Review*, Vol. 54, No. 3, 1937–1967. doi:10.1007/s10462-020-09896-5.

88. Pourtaheri, Z. K., and Zahiri, S. H. (2016). Ensemble classifiers with improved overfitting, *2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC)*, IEEE, 93–97. doi:10.1109/CSIEC.2016.7482130.

89. Tian, Y., and Zhang, Y. (2022). A comprehensive survey on regularization strategies in machine learning, *Information Fusion*, Vol. 80, 146–166. doi:10.1016/j.inffus.2021.11.005.

90. Berrar, D. (2019). Cross-Validation.

91. Handa, K., Sakamoto, S., Kageyama, M., and Iijima, T. (2023). Development of a 2D-QSAR Model for Tissue-to-Plasma Partition Coefficient Value with High Accuracy Using Machine Learning Method, Minimum Required Experimental Values, and Physicochemical Descriptors, *European Journal of Drug Metabolism and Pharmacokinetics*, Vol. 48, No. 4, 341–352. doi:10.1007/s13318-023-00832-w.

92. Ramaneswaran, S., Srinivasan, K., Vincent, P. M. D. R., and Chang, C.-Y. (2021). Hybrid Inception v3 XGBoost Model for Acute Lymphoblastic Leukemia Classification, *Computational and Mathematical Methods in Medicine*, Vol. 2021, 1–10. doi:10.1155/2021/2577375.

93. Le, T.-T.-H., Kim, H., Kang, H., and Kim, H. (2022). Classification and Explanation for Intrusion Detection System Based on Ensemble Trees and SHAP Method, *Sensors*, Vol. 22, No. 3, 1154. doi:10.3390/s22031154.

94. Ekanayake, I. U., Palitha, S., Gamage, S., Meddage, D. P. P., Wijesooriya, K., and Mohotti, D. (2023). Predicting adhesion strength of micropatterned surfaces using gradient boosting models and explainable artificial intelligence visualizations, *Materials Today Communications*, Vol. 36, 106545. doi:10.1016/j.mtcomm.2023.106545.

95. Tian, H., You, S., Xiong, T., Ji, M., Zhang, K., Jiang, L., Du, T., Li, Y., Liu, W., and Lin, S. (2023). Discovery of a Novel Photocaged PI3K Inhibitor Capable of Real-Time Reporting of Drug Release, *ACS Medicinal Chemistry Letters*, Vol. 14, No. 8, 1100–1107.

96. Zhang, S., Bamakan, S. M. H., Qu, Q., and Li, S. (2018). Learning for personalized medicine: a comprehensive review from a deep learning perspective, *IEEE Reviews in Biomedical Engineering*, Vol. 12, 194–208.