FUTURE TECHNO SCIENCE

Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

An Interpretable Machine Learning Strategy for Antimalarial Drug Discovery with LightGBM and SHAP

Teuku Rizky Noviandy *, Ghalieb Mutig Idroes, and Irsan Hardi

Interdisciplinary Innovation Research Unit, Graha Primera Saintifika, Aceh Besar 23771, Indonesia; email: trizkynoviandy@gmail.com, ghaliebidroes@outlook.com, irsan.hardi@gmail.com

* Corresponding Author : Teuku Rizky Noviandy

Abstract: Malaria continues to pose a significant global health threat, and the emergence of drug-resistant malaria exacerbates the challenge, underscoring the urgent need for new antimalarial drugs. While several machine learning algorithms have been applied to quantitative structure-activity relationship (QSAR) modeling for antimalarial compounds, there remains a need for more interpretable models that can provide insights into the underlying mechanisms of drug action, facilitating the rational design of new compounds. This study develops a QSAR model using Light Gradient Boosting Machine (LightGBM). The model is integrated with SHapley Additive exPlanations (SHAP) to enhance interpretability. The LightGBM model demonstrated superior performance in predicting antimalarial activity, with an ac-curacy of 86%, precision of 85%, sensitivity of 81%, specificity of 89%, and an F1-score of 83%. SHAP analysis identified key molecular descriptors such as maxdO and GATS2m as significant contributors to antimalarial activity. The integration of LightGBM with SHAP not only enhances the predictive ac-curacy of the QSAR model but also provides valuable insights into the importance of features, aiding in the rational design of new antimalarial drugs. This approach bridges the gap between model accuracy and interpretability, offering a robust framework for efficient and effective drug discovery against drug-resistant malaria strains.

Keywords: Classification; Gradient boosting; Molecular descriptor; QSAR; Supervised learning

1. Introduction

Malaria is a serious infectious disease that affects red blood cells. It is caused by the parasite *Plasmodium falciparum* and is transmitted through the bite of female *Anopheles* mosquitoes [1]. Between 2000 and 2020, an estimated 1.7 billion malaria cases and 10.6 million malaria-related deaths occurred globally [2]. Reports from the Centers for Disease Control and Prevention (CDC) indicate the emergence of drug-resistant malaria cases, posing a significant threat to malaria control and leading to increased morbidity and mortality rates [3]. The high incidence of malaria and the occurrence of drug resistance highlight the urgent need for new drug candidates [4].

High-throughput screening (HTS) is a commonly used method to discover new drug candidates; however, it is time-consuming, costly, and has a low hit rate [5], [6]. An alternative approach to discovering new drug compounds is using quantitative structure-activity relationship (QSAR) models. These models leverage machine learning and statistical techniques to identify the relationship between the chemical structure of a compound and its biological activity [7]–[9]. By training algorithms on datasets of known compounds and their activities, QSAR models can predict the efficacy of new compounds. This in-silico method allows researchers to computationally screen and prioritize candidate compounds with desired biological activities, significantly reducing the need for extensive in vivo testing [10], [11]. The QSAR-based selection process is highly efficient, offering rapid and accurate predictions with a high hit rate [12].

In recent years, machine learning has revolutionized the QSAR methodology, offering enhanced capabilities in predicting the biological activities of compounds based on their

Received: July, 9th 2024
Revised: August, 5th 2024
Accepted: August, 6th 2024
Published: August, 7th 2024
Curr. Ver.: August, 7th 2024

DOI : 10.62411/faith.2024-16

chemical structures [13], [14]. Advanced machine learning algorithms, such as random forests, support vector machines, and deep learning models, have demonstrated superior performance in identifying potential drug candidates with higher accuracy and efficiency than traditional QSAR approaches [15]–[17]. These algorithms can handle large datasets, uncover complex patterns, and improve the predictive power of QSAR models, thereby accelerating the drug discovery process and addressing the challenges posed by drug-resistant malaria.

One popular method for building robust QSAR models is gradient boosting, which has gained significant traction due to its high predictive performance and efficiency. Among various gradient boosting frameworks, LightGBM (Light Gradient Boosting Machine) stands out for its speed and accuracy [18], [19]. LightGBM is designed to handle large-scale data with lower memory usage and faster training times, making it particularly suitable for high-throughput QSAR applications. Its effectiveness in capturing intricate patterns within the data and its scalability makes it an excellent choice for drug discovery [20].

An interpretable approach in machine learning for QSAR is particularly valuable, as it allows researchers to understand the underlying mechanisms driving the predictions and make informed decisions [21]. Interpretable models provide insights into which chemical features are most influential in determining biological activity, facilitating the design of new compounds with optimized properties [22]. This transparency enhances trust in the model's predictions and aids in the rational design of drugs, ultimately leading to more effective and targeted malaria treatments.

The primary aim of this study is to develop a robust and interpretable machine learning-based QSAR model to predict the antimalarial activity of chemical compounds. By integrating state-of-the-art machine learning techniques focusing on interpretability, this study seeks to identify novel drug candidates that are effective against *Plasmodium falciparum*, including drug-resistant strains. Additionally, the study aims to provide a framework for understanding the relationship between chemical structures and their antimalarial properties, thereby contributing to the rational design of new antimalarial drugs.

The contributions of this study are as follows:

- We developed an advanced QSAR model using LightGBM, focusing on interpretability to provide clear insights into the model's decision-making process.
- Through SHAP analysis, we identified critical molecular descriptors that significantly contribute to antimalarial activity, aiding in the rational design of new compounds.
- Our model successfully predicted potential new drug candidates that are effective against *Plasmodium falciparum*.
- We established a robust framework for integrating machine learning and QSAR methodologies in drug discovery, enhancing the efficiency and accuracy of identifying promising drug candidates.

This paper is structured as follows: Section 2 reviews related works, providing an overview of existing QSAR methodologies and the application of machine learning in drug discovery, with a particular focus on antimalarial compounds. Section 3 details the methodology used in this study, including data collection, feature extraction, model development using LightGBM, and the interpretability techniques employed. Section 4 presents the results and discussion, showcasing the performance of the developed QSAR model, key findings, and their implications for antimalarial drug discovery. Finally, Section 5 concludes the paper, summarizing the main contributions and highlighting the study's potential impact.

2. Related Work

The integration of machine learning into QSAR modeling has significantly advanced the field of computational drug discovery. Traditional QSAR methods relied on linear regression and simple statistical techniques and often fell short in capturing the complex relationships between chemical structures and their biological activities [23]. The advent of machine learning has addressed these limitations by introducing non-linear models capable of handling high-dimensional data and uncovering intricate patterns.

The application of machine learning-based QSAR models in malaria research has shown promising results in identifying potential antimalarial compounds. Given the urgent need for new treatments due to the rise of drug-resistant strains of *Plasmodium falciparum*, researchers have increasingly turned to advanced computational techniques. Studies have utilized various machine learning algorithms to build QSAR models that predict the efficacy of chemical compounds against malaria. For instance, Azmi et al. [24] utilized genetic algorithms and artificial neural networks (ANN) to predict the activity of 61 fusidic acid compounds as antimalarial agents, achieving accuracies of 0.96 and 0.92 based on training and test data evaluations, respectively. Similarly, Egieyeh et al. [25] classified 1155 antimalarial compounds using naïve Bayes, support vector machine (SVM), and voted perceptron methods, with SVM yielding the highest accuracy of 0.85 on test data. Additionally, Danishuddin [26] employed several methods including XGBoost, SVM, k-Nearest Neighbors (KNN), and random forest to predict 4750 antimalarial compounds, demonstrating that XGBoost achieved high classification performance with an accuracy of 0.86 and an AUC of 0.91 on test data. Furthermore, Mswahili et al. developed five machine learning models to predict antimalarial bioactivities of a drug against *Plasmodium falciparum*, with XGBoost achieving an accuracy of 83% [27].

Moreover, studies have explored combining machine learning with traditional cheminformatics techniques to enhance the interpretability of QSAR models. Daoui et al. [28] integrated machine learning models with molecular docking simulations and absorption, distribution, metabolism, excretion and toxicity (ADMET) properties in silico studies to predict potent anti-tumor agents, providing interpretable insights into compound efficacy and safety. Ashraf et al. [29] combined machine learning with 3D QSAR, molecular docking, and dynamics simulations studies to model and design TTK inhibitors, offering an interpretable framework for understanding the molecular interactions and dynamics of potential drug candidates.

Despite these advancements, there remains a need for the development of more sophisticated and interpretable models. This would not only enhance the predictive accuracy but also provide deeper insights into the underlying mechanisms of drug action, thereby facilitating the discovery of novel antimalarial compounds.

3. Proposed Method

3.1. Dataset

This study utilized a dataset obtained from the research conducted by Danishuddin et al. [26]. The dataset consists of 4750 compound samples, each labeled as active or inactive. Each compound is represented by 98 molecular descriptors, which are quantitative representations of molecular properties [30], used as features for model construction. This dataset was selected due to its comprehensive representation of molecular properties and the balanced nature of its labels, making it highly suitable for building and evaluating predictive models. Detailed information about the dataset can be found in the original paper by Danishuddin et al. [26].

Figure 1 depicts the visualization of compound distribution with active and inactive classes using principal component analysis (PCA). It can be observed that the data cannot be linearly separated, and samples from both classes overlap and are difficult to distinguish, posing a challenge in the model training process.

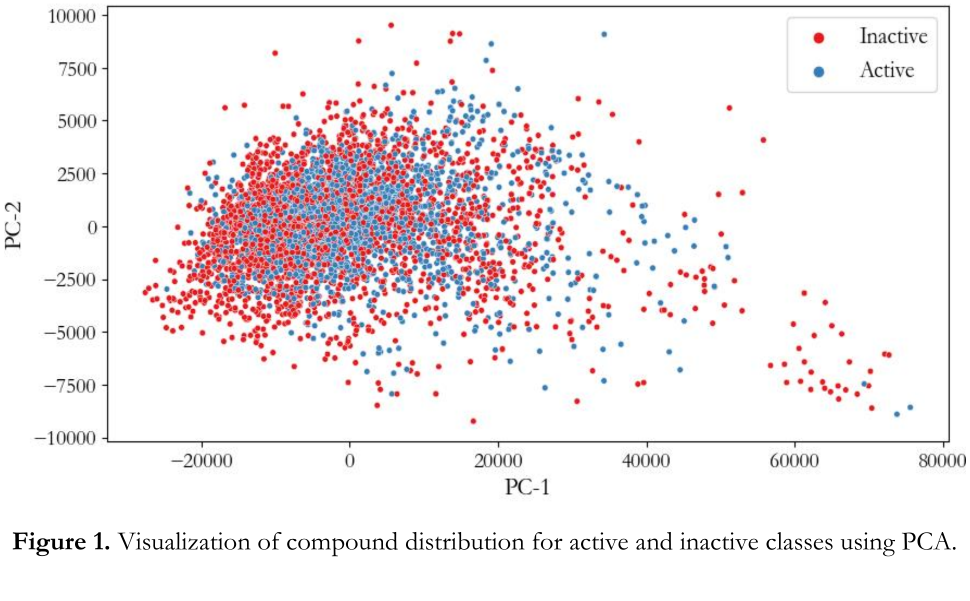

Figure 1. Visualization of compound distribution for active and inactive classes using PCA.

3.2. Data Preprocessing

For preprocessing, we standardized the molecular descriptors by removing the mean and scaling to unit variance to ensure that all features contribute equally to the model and to improve the convergence of gradient-based optimization algorithms [31], [32]. The active and inactive labels were encoded into numerical values. Then, the dataset was divided into two subsets: 80% for training data and 20% for testing data [33]. The class distribution in each subset is presented in Table 1.

Table 1. Hyperparameter search space and selected values for tuning.

Subset	Class	Samples	Percentages (%)
Training	Active	1673	43.04
	Inactive	2214	56.96
Testing	Active	375	43.45
	Inactive	488	56.55

3.3. Model Training

This study proposes a machine learning-based QSAR model utilizing LightGBM for its high predictive performance and efficiency. LightGBM is particularly suited for high-dimensional datasets due to its fast-training speed and low memory usage [34], [35]. We chose LightGBM over deep learning models such as LSTM and GRU because, although deep learning models are powerful, they are typically more computationally intensive and require larger datasets to achieve optimal performance. Recent studies have also found that gradient boosting models perform better than deep learning models for tabular machine learning tasks. The overall workflow of our proposed approach, including data preprocessing, model training, hyperparameter optimization, and interpretability analysis, is visualized in Figure 2.

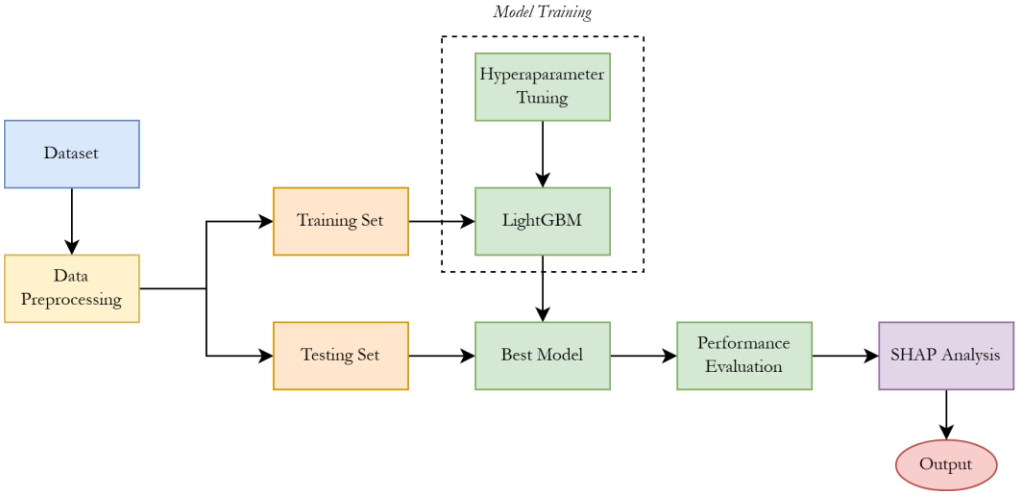

Figure 2. Workflow of the proposed approach.

The objective function optimized by LightGBM for binary classification is the binary log loss, defined as shown in Equation 1:

$$L(y,\hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{1}$$

where y_i is the true label, $\hat{y}_i$ is the predicted probability, and N is the number of samples.

To optimize the performance of the LightGBM model, we employ Optuna, an automatic hyperparameter optimization software framework. Optuna uses a Bayesian optimization approach to efficiently search for the best hyperparameter values [36]. This method balances exploration and exploitation, allowing for a more efficient search compared to traditional methods. Optuna’s versatility in defining complex search spaces and its support for various samplers make it a powerful tool for hyperparameter tuning [37].

The specific hyperparameter search space and the selected values for tuning are outlined in Table 2. This table details the range of values considered for each hyperparameter and the optimal values determined through the tuning process.

Table 2. Hyperparameter search space and selected values for tuning.

Hyperparameter	Search Space
num_leaves	Integer between 2 and 256
learning_rate	Log-uniform distribution between 0.005 and 0.5
feature_fraction	Uniform distribution between 0.1 and 1.0
bagging_fraction	Uniform distribution between 0.1 and 1.0
bagging_freq	Integer between 1 and 10
min_child_samples	Integer between 5 and 100

3.4. Performance Evaluation

To evaluate the performance of our proposed LightGBM-based QSAR model, we utilize several common metrics in binary classification: accuracy, precision, sensitivity (recall), specificity, and F1-score [38]–[40]. These metrics provide a comprehensive assessment of the model's predictive capabilities, particularly to identify active and inactive compounds:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{6}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives [41].

In addition to evaluating the LightGBM model, we also compared our approach with other machine learning methods, including XGBoost, Random Forest, k-Nearest Neighbors (KNN), and Logistic Regression [42]–[44]. These comparisons will help benchmark our proposed model's performance against established algorithms and provide a comprehensive understanding of its strengths and weaknesses in predicting antimalarial activity.

3.5. SHAP Analysis

To enhance the interpretability of our model, we integrate SHAP (SHapley Additive exPlanations) values, which provide insights into the contributions of individual features to the model's predictions [45], [46]. SHAP provide a unified measure of feature importance. SHAP values are derived from cooperative game theory and provide insights into the contribution of each feature to the model's predictions [47], [48].

Due to several key advantages, we selected SHAP over other interpretability methods such as Local Interpretable Model-agnostic Explanations (LIME), Permutation Feature Importance, and Partial Dependence Plots (PDP). SHAP provides a consistent and theoretically sound approach to feature attribution by ensuring that the contributions of features are fairly distributed according to their marginal contributions across all possible feature subsets [49]. This property makes SHAP particularly robust for capturing the interactions between features and their impact on predictions [50]. In contrast, LIME provides local approximations of model behavior but may lack consistency across different local regions of the feature space.

Permutation Feature Importance and PDP offer valuable insights but can be less robust in capturing feature interactions and may not be as effective in explaining complex models comprehensively [51], [52].

The integration of LightGBM with SHAP not only enhances the predictive power of our QSAR model but also ensures that the model remains interpretable, allowing researchers to make informed decisions based on the model's output. This approach bridges the gap between model accuracy and interpretability, providing a robust framework for the discovery of new antimalarial drug candidates.

4. Results and Discussion

The hyperparameter optimization process using Optuna yielded the best set of hyperparameters for the LightGBM model. The optimal values identified were as follows: the number of leaves was set to 175, the learning rate was 0.104, the feature fraction was 0.806, the bagging fraction was 0.978, the bagging frequency was 5, and the minimum number of child samples was 17. These parameters were used to train the final QSAR model for predicting the antimalarial activity of chemical compounds.

Table 3. Performance comparison of machine learning models.

Model	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
LightGBM	86	85	81	89	83
XGBoost	83	83	78	88	80
Random Forest	83	86	72	91	78
KNN	65	61	58	72	59
Logistic Regression	69	70	49	84	57

Table 3 presents the performance metrics for the LightGBM model, showcasing its clear superiority over other machine learning models. The LightGBM model achieved an accuracy of 86%, significantly higher than XGBoost (83%), Random Forest (83%), KNN (65%), and Logistic Regression (69%), indicating its exceptional ability to classify active and inactive compounds correctly. The model's precision is 85%, meaning a high proportion of true positive predictions, which is comparable to Random Forest's 86% but better than XGBoost (83%), KNN (61%), and Logistic Regression (70%).

The sensitivity of the LightGBM model is 81%, superior to XGBoost (78%), Random Forest (72%), KNN (58%), and Logistic Regression (49%). This high sensitivity ensures that active compounds are effectively identified, reducing the risk of false negatives. The LightGBM model also demonstrated a high specificity of 89%, slightly lower than Random Forest (91%) but higher than XGBoost (88%), KNN (72%), and Logistic Regression (84%), which is crucial for minimizing false positives. The F1-score, which balances precision and recall, was highest for the LightGBM model at 83%, surpassing XGBoost (80%), Random Forest (78%), KNN (59%), and Logistic Regression (57%), further highlighting its overall effectiveness.

The superior performance of the LightGBM model can be attributed to its ability to handle the nonlinearity of data. LightGBM's gradient boosting framework effectively combines multiple weak learners to form a strong predictive model, allowing it to manage high-dimensional data and complex relationships between chemical structures and their biological activities. In contrast, models like KNN and Logistic Regression struggled due to these complexities. KNN, with its simplistic approach of classifying based on nearest neighbors, failed to capture the intricate patterns, resulting in low accuracy (65%) and high misclassification rates. Similarly, Logistic Regression, a linear model, could not handle the nonlinear relationships within the dataset, leading to the lowest sensitivity (49%) and a high number of false positives and false negatives.

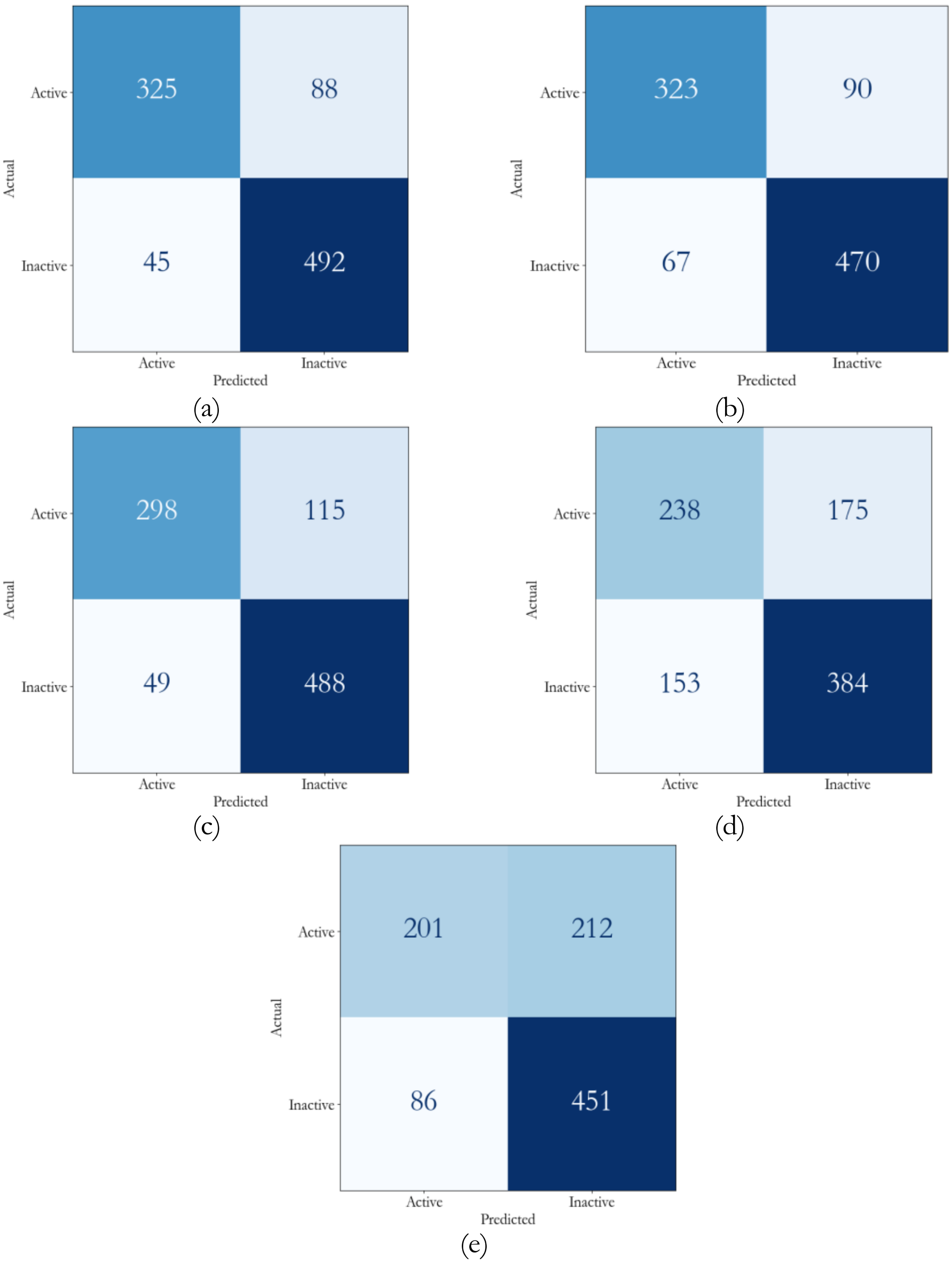

Figure 3. Confusion matrix (**a**) LightGBM, (**b**) XGBoost, (**c**) Random Forest, (**d**) KNN, (**e**) Logistic Regression.

The confusion matrices reveal valuable insights into the performance of different models in classifying active and inactive compounds. The LightGBM model, depicted in Figure 3a, stands out with 325 true positives and 492 true negatives, showcasing its high accuracy in correctly identifying compounds. Its low numbers of false positives (88) and false negatives (45) further emphasize its robustness and ability to minimize errors. In comparison, the XGBoost model (Figure 3b) shows good performance but has slightly more false positives (90) and false negatives (67), making it less accurate than LightGBM and more likely to miss active compounds. The Random Forest model (Figure 3c) has a similar true negative count (488) but higher false positives (115) and slightly fewer false negatives (49), indicating a more conservative approach with increased false positives. The KNN model, shown in Figure 3d, exhibits significant misclassification with high false positives (175) and false negatives (153), proving less effective for this task due to its sensitivity to overlapping data and inability to capture complex patterns. Logistic Regression (Figure 3e) also struggles, with 201 true positives and 451 true negatives, and the highest false positives (212) and substantial false negatives (86), indicating its inadequacy in handling the dataset's complexity. The LightGBM model's superior performance in minimizing misclassification makes it the most effective model for predicting antimalarial activity in this study.

Figure 4 shows the Receiver Operating Characteristic (ROC) curves for the different machine learning models and provide insights into their performance in predicting the antimalarial activity of chemical compounds by plotting the true positive rate (sensitivity) against the false positive rate (1 - specificity) for various thresholds. The ROC curve for the LightGBM model (blue line) demonstrates superior performance with the highest area under the curve (AUC) value of 0.92, indicating its excellent ability to discriminate between active and inactive compounds. XGBoost (orange line) follows closely with an AUC of 0.91, showing strong performance but slightly less than LightGBM. The Random Forest model (green line) also has an AUC of 0.91, suggesting comparable effectiveness to XGBoost. In contrast, the KNN model (red line) exhibits a significantly lower AUC of 0.67, reflecting poor performance and a higher misclassification rate. Logistic Regression (purple line) has an AUC of 0.70, which is better than KNN but still considerably lower than LightGBM, XGBoost, and Random Forest, indicating moderate performance and limitations in capturing dataset complexities.

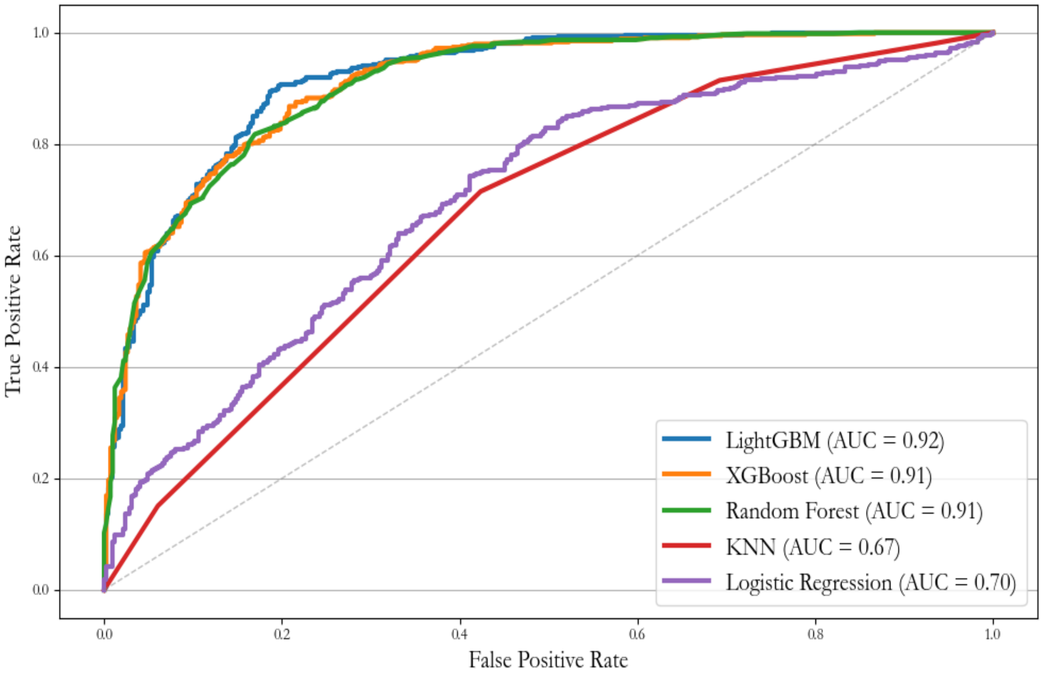

Figure 4. ROC curves for different machine learning models.

Furthermore, we performed SHAP analysis to interpret the LightGBM's predictions and identify the most influential molecular descriptors. Figure 5 illustrates the top 10 most important molecular descriptors based on their mean absolute SHAP values, which indicate their average impact on the model's output. The most important descriptor is maxdO, which has the highest SHAP value, suggesting it has the greatest influence on the model's predictions. This descriptor likely captures a critical structural or chemical property relevant to antimalarial activity. The second most important descriptor, GATS2m, also significantly impacts the model's predictions, albeit to a lesser extent than maxdO. This descriptor might be associated with specific molecular interactions or properties crucial for distinguishing active compounds. While still influential, the tenth most important descriptor, SpMax5_Bhs, has a smaller SHAP value than the top descriptors, indicating a relatively lower but still meaningful contribution to the model's decision-making process. The combined insights from these descriptors help us understand the underlying factors driving the LightGBM model's high predictive accuracy.

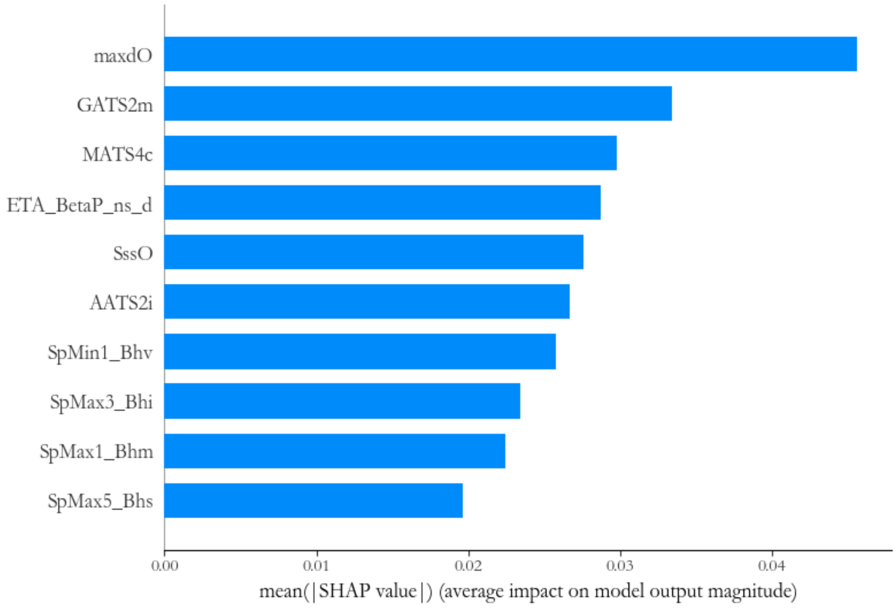

Figure 5. Top 10 most important molecular descriptors based on their mean absolute SHAP values using the LightGBM model.

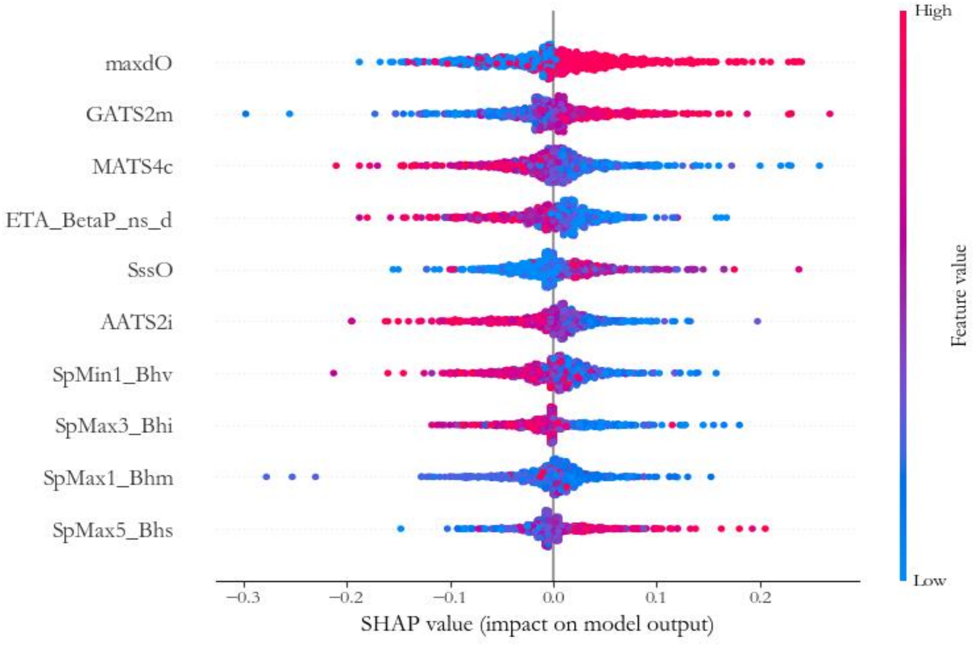

Figure 6. SHAP beeswarm plot showing the importance and impact of various molecular descriptors in predicting the antimalarial activity of compounds using the LightGBM model.

Figure 6 presents a SHAP beeswarm plot, which provides a detailed visualization of various molecular descriptors' importance and impact in predicting compounds' antimalarial activity using the LightGBM model. Each dot in the plot represents a SHAP value for a specific feature, color-coded by the feature's value (with high values in red and low values in blue). This type of plot allows for a nuanced understanding of how each feature affects individual predictions. For example, the feature maxdO shows the highest SHAP values, indicating it significantly influences the model's predictions, with higher values (red) generally increasing the predicted antimalarial activity. Conversely, GATS2m has a more balanced distribution of high and low values, suggesting a more complex relationship with the model's output. Other features like MATS4c, ETA_BetaP_ns_d, and SssO also demonstrate significant impacts, each with distinct influence patterns. The beeswarm plot's ability to show the distribution and variability of feature impacts, as well as the interaction between different feature values, complements the bar plot shown in Figure 5, which ranks features based on their average absolute SHAP values. While the bar plot provides a clear and straightforward ranking of feature importance, the beeswarm plot offers deeper insights into features' contextual and interaction effects on model predictions.

To further validate our findings, we compared our results with a previous study by Danishuddin et al. and found that while both studies achieved the same accuracy (86%), our study achieved a higher AUC (92%) compared to the previous AUC (91%). Although the improvement in AUC is not substantial, it indicates a better overall model performance in distinguishing between active and inactive compounds. Additionally, our LightGBM model is interpretable with SHAP, providing insights into the contribution of each feature to the model's predictions, which is crucial for understanding the underlying mechanisms of antimalarial activity.

The results of this study demonstrate the effectiveness of integrating LightGBM with SHAP for developing an interpretable QSAR model to predict the antimalarial activity of chemical compounds. The superior performance of the LightGBM model, as evidenced by its high accuracy, precision, sensitivity, specificity, and F1-score, underscores its robustness in handling high-dimensional data and complex relationships between molecular descriptors and biological activity. The use of SHAP values provides valuable insights into the importance of features and the contribution of individual descriptors.

However, this approach has several limitations. First, while SHAP enhances interpretability, it does not inherently address model biases or data quality issues, which can affect predictions. Second, relying on a static dataset limits the model's ability to generalize to new, unseen compounds, highlighting the need for continuous model validation and updating with diverse and larger datasets. Third, the computational complexity of hyperparameter tuning with Optuna and interpreting SHAP values can be resource-intensive, necessitating robust computational infrastructure and expertise. Future work should focus on addressing these limitations by exploring transfer learning techniques, expanding the dataset with new compounds, and optimizing computational resources to streamline model training and interpretation processes. This would further enhance the model's applicability and reliability in the drug discovery process, paving the way for more efficient and effective identification of antimalarial candidates.

6. Conclusions

This study presents a robust and interpretable machine learning-based QSAR model using LightGBM to predict the antimalarial activity of chemical compounds against *Plasmodium falciparum*. The LightGBM model achieved an accuracy of 86%, precision of 85%, sensitivity of 81%, specificity of 89%, and an F1-score of 83%, significantly outperforming other models like XGBoost and Random Forest. By integrating SHAP values, we provided a detailed understanding of the importance of features, aiding in the rational design of new antimalarial drugs. This study establishes a robust framework for integrating machine learning with QSAR methodologies, bridging the gap between predictive accuracy and interpretability. This approach not only accelerates the drug discovery process but also provides a clear pathway for developing new antimalarial drugs, including those targeting drug-resistant malaria strains.

Author Contributions: Conceptualization: T.R.N. and G.M.I.; methodology, T.R.N.; software: T.R.N.; validation: T.R.N. and I.H..; formal analysis: T.R.N. and G.M.I.; investigation: T.R.N. and G.M.I.; resources: T.R.N. and I.H.; data curation: G.M.I. and I.H.; writing—original draft preparation: T.R.N. and I.H.; writing—review and editing: T.R.N. and G.M.I.; visualization: G.M.I.; supervision: T.R.N.; project administration: T.R.N. and I.H.; funding acquisition: T.R.N.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study was retrieved from the research conducted by Danishuddin et al., available at https://doi.org/10.1080/1062936X.2019.1635526.

Acknowledgments: We would like to thank Graha Primera Saintifika for their assistance and support in this research.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] E. Scholar, “Malaria,” in *xPharm: The Comprehensive Pharmacology Reference*, Elsevier, 2007, pp. 1–5. doi: 10.1016/B978-008055232-3.60932-8.

[2] World Health Organization, *World malaria report 2021*. World Health Organization, 2021.

[3] CDC, “Drug Resistance in the Malaria-Endemic World,” 2018.

[4] M. M. Ippolito, K. A. Moser, J.-B. B. Kabuya, C. Cunningham, and J. J. Juliano, “Antimalarial Drug Resistance and Implications for the WHO Global Technical Strategy,” *Curr. Epidemiol. Reports*, vol. 8, no. 2, pp. 46–62, Mar. 2021, doi: 10.1007/s40471-021-00266-5.

[5] B. J. Neves, R. C. Braga, C. C. Melo-Filho, J. T. Moreira-Filho, E. N. Muratov, and C. H. Andrade, “QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery,” *Front. Pharmacol.*, vol. 9, Nov. 2018, doi: 10.3389/fphar.2018.01275.

[6] M. Abdullahi, G. A. Shallangwa, and A. Uzairu, “In silico QSAR and molecular docking simulation of some novel aryl sulfonamide derivatives as inhibitors of H5N1 influenza A virus subtype,” *Beni-Suef Univ. J. Basic Appl. Sci.*, vol. 9, no. 1, p. 2, Dec. 2020, doi: 10.1186/s43088-019-0023-y.

[7] S. A. Alsenan, I. M. Al-Turaiki, and A. M. Hafez, “Feature Extraction Methods in Quantitative Structure–Activity Relationship Modeling: A Comparative Study,” *IEEE Access*, vol. 8, pp. 78737–78752, 2020, doi: 10.1109/ACCESS.2020.2990375.

[8] T. R. Noviandy, K. Nisa, G. M. Idroes, I. Hardi, and N. R. Sasmita, “Classifying Beta-Secretase 1 Inhibitor Activity for Alzheimer’s Drug Discovery with LightGBM,” *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 358–367, Mar. 2024, doi: 10.62411/jcta.10129.

[9] P. Carracedo-Reboredo *et al.*, “A review on machine learning approaches and trends in drug discovery,” *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 4538–4558, 2021, doi: 10.1016/j.csbj.2021.08.011.

[10] S. Kwon, H. Bae, J. Jo, and S. Yoon, “Comprehensive ensemble in QSAR prediction for drug discovery,” *BMC Bioinformatics*, vol. 20, no. 1, p. 521, Dec. 2019, doi: 10.1186/s12859-019-3135-4.

[11] K. B. Jillahi and A. Iorliam, “A Scoping Literature Review of Artificial Intelligence in Epidemiology: Uses, Applications, Challenges and Future Trends,” *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 421–445, Apr. 2024, doi: 10.62411/jcta.10350.

[12] P. G. R. Achary, “Applications of Quantitative Structure-Activity Relationships (QSAR) based Virtual Screening in Drug Design: A Review,” *Mini-Reviews Med. Chem.*, vol. 20, no. 14, pp. 1375–1388, Sep. 2020, doi: 10.2174/1389557520666200429102334.

[13] L. Patel, T. Shukla, X. Huang, D. W. Ussery, and S. Wang, “Machine Learning Methods in Drug Discovery,” *Molecules*, vol. 25, no. 22, p. 5277, Nov. 2020, doi: 10.3390/molecules25225277.

[14] H. Li, K. Sze, G. Lu, and P. J. Ballester, “Machine‐learning scoring functions for structure‐based drug lead optimization,” *WIREs Comput. Mol. Sci.*, vol. 10, no. 5, Sep. 2020, doi: 10.1002/wcms.1465.

[15] T. R. Noviandy, A. Maulana, T. Bin Emran, G. M. Idroes, and R. Idroes, “QSAR Classification of Beta-Secretase 1 Inhibitor Activity in Alzheimer’s Disease Using Ensemble Machine Learning Algorithms,” *Heca J. Appl. Sci.*, vol. 1, no. 1, pp. 1–7, May 2023, doi: 10.60084/hjas.v1i1.12.

[16] F. Rahman, K. M. Lhaksmana, and I. Kurniawan, “Implementation of Simulated Annealing-Support Vector Machine on QSAR Study of Fusidic Acid Derivatives as Anti-Malarial Agent,” in *2020 6th International Conference on Interactive Digital Media (ICIDM)*, Dec. 2020, pp. 1–4. doi: 10.1109/ICIDM51048.2020.9339632.

[17] Y. Matsuzaka, T. Hosaka, A. Ogaito, K. Yoshinari, and Y. Uesawa, “Prediction Model of Aryl Hydrocarbon Receptor Activation by a Novel QSAR Approach, DeepSnap–Deep Learning,” *Molecules*, vol. 25, no. 6, p. 1317, Mar. 2020, doi: 10.3390/molecules25061317.

[18] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[19] T. R. Noviandy *et al.*, “Integrating Genetic Algorithm and LightGBM for QSAR Modeling of Acetylcholinesterase Inhibitors in Alzheimer’s Disease Drug Discovery,” *Malacca Pharm.*, vol. 1, no. 2, pp. 48–54, Jul. 2023, doi: 10.60084/mp.v1i2.60.

[20] L. Patel, T. Shukla, X. Huang, D. W. Ussery, and S. Wang, “Machine Learning Methods in Drug Discovery,” *Molecules*, vol. 25, no. 22, p. 5277, Nov. 2020, doi: 10.3390/molecules25225277.

[21] R. Dybowski, “Interpretable machine learning as a tool for scientific discovery in chemistry,” *New J. Chem.*, vol. 44, no. 48, pp. 20914–20920, 2020, doi: 10.1039/D0NJ02592E.

[22] T. R. Noviandy, G. M. Idroes, and I. Hardi, “Machine Learning Approach to Predict AXL Kinase Inhibitor Activity for Cancer Drug Discovery Using XGBoost and Bayesian Optimization,” *J. Soft Comput. Data Min.*, vol. 5, no. 1, pp. 46–56, Jun. 2024.

[23] T. Puzyn, J. Leszczynski, and M. T. Cronin, *Recent Advances in QSAR Studies*, vol. 8. Dordrecht: Springer Netherlands, 2010. doi: 10.1007/978-1-4020-9783-6.

[24] H. F. Azmi, K. M. Lhaksmana, and I. Kurniawan, “QSAR Study of Fusidic Acid Derivative as Anti-Malaria Agents by using Artificial Neural Network-Genetic Algorithm,” in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, Jun. 2020, pp. 1–4. doi: 10.1109/ICoICT49345.2020.9166158.

[25] S. Egieyeh, J. Syce, S. F. Malan, and A. Christoffels, “Predictive classifier models built from natural products with antimalarial bioactivity using machine learning approach,” *PLoS One*, vol. 13, no. 9, p. e0204644, Sep. 2018, doi: 10.1371/journal.pone.0204644.

[26] Danishuddin, G. Madhukar, M. Z. Malik, and N. Subbarao, “Development and rigorous validation of antimalarial predictive models using machine learning approaches,” *SAR QSAR Environ. Res.*, vol. 30, no. 8, pp. 543–560, Aug. 2019, doi: 10.1080/1062936X.2019.1635526.

[27] M. E. Mswahili, G. L. Martin, J. Woo, G. J. Choi, and Y.-S. Jeong, “Antimalarial Drug Predictions Using Molecular Descriptors and Machine Learning against Plasmodium Falciparum,” *Biomolecules*, vol. 11, no. 12, p. 1750, Nov. 2021, doi: 10.3390/biom11121750.

[28] O. Daoui, S. Elkhattabi, S. Chtita, R. Elkhalabi, H. Zgou, and A. T. Benjelloun, "QSAR, molecular docking and ADMET properties in silico studies of novel 4,5,6,7-tetrahydrobenzo[D]-thiazol-2-Yl derivatives derived from dimedone as potent anti-tumor agents through inhibition of C-Met receptor tyrosine kinase," *Heliyon*, vol. 7, no. 7, p. e07463, Jul. 2021, doi: 10.1016/j.heliyon.2021.e07463.

[29] N. Ashraf *et al.*, "Combined 3D-QSAR, molecular docking and dynamics simulations studies to model and design TTK inhibitors," *Front. Chem.*, vol. 10, Nov. 2022, doi: 10.3389/fchem.2022.1003816.

[30] R. Idroes *et al.*, "Application of Genetic Algorithm-Multiple Linear Regression and Artificial Neural Network Determinations for Prediction of Kovats Retention Index," *Int. Rev. Model. Simulations*, vol. 14, no. 2, p. 137, Apr. 2021, doi: 10.15866/iremos.v14i2.20460.

[31] G. M. Idroes, I. Hardi, I. S. Hilal, R. T. Utami, T. R. Noviandy, and R. Idroes, "Economic Growth and Environmental Impact: Assessing the Role of Geothermal Energy in Developing and Developed Countries," *Innov. Green Dev.*, vol. 3, no. 3, p. 100144, Sep. 2024, doi: 10.1016/j.igd.2024.100144.

[32] G. M. Idroes, I. Hardi, M. H. Rahman, M. Afjal, T. R. Noviandy, and R. Idroes, "The Dynamic Impact of Non-renewable and Renewable Energy on Carbon Dioxide Emissions and Ecological Footprint in Indonesia," *Carbon Res.*, vol. 3, no. 1, p. 35, Apr. 2024, doi: 10.1007/s44246-024-00117-0.

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

[34] T. R. Noviandy, S. I. Nainggolan, R. Raihan, I. Firmansyah, and R. Idroes, "Maternal Health Risk Detection Using Light Gradient Boosting Machine Approach," *Infolitika J. Data Sci.*, vol. 1, no. 2, pp. 48–55, Dec. 2023, doi: 10.60084/ijds.v1i2.123.

[35] H. Yang, Z. Chen, H. Yang, and M. Tian, "Predicting Coronary Heart Disease Using an Improved LightGBM Model: Performance Analysis and Comparison," *IEEE Access*, vol. 11, pp. 23366–23380, 2023, doi: 10.1109/ACCESS.2023.3253885.

[36] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[37] S. Shekhar, A. Bansode, and A. Salim, "A Comparative study of Hyper-Parameter Optimization Tools," in *2021 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*, Dec. 2021, pp. 1–6. doi: 10.1109/CSDE53843.2021.9718485.

[38] A. N. Safriandono, D. R. I. M. Setiadi, A. Dahlan, F. Z. Rahmanti, I. S. Wibisono, and A. A. Ojugo, "Analyzing Quantum Feature Engineering and Balancing Strategies Effect on Liver Disease Classification," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 51–63, Jun. 2024, doi: 10.62411/faith.2024-12.

[39] G. M. Idroes *et al.*, "Urban Air Quality Classification Using Machine Learning Approach to Enhance Environmental Monitoring," *Leuser J. Environ. Stud.*, vol. 1, no. 2, pp. 62–68, Nov. 2023, doi: 10.60084/ljes.v1i2.99.

[40] D. R. I. M. Setiadi, H. M. M. Islam, G. A. Trisnapradika, and W. Herowati, "Analyzing Preprocessing Impact on Machine Learning Classifiers for Cryotherapy and Immunotherapy Dataset," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 39–50, Jun. 2024, doi: 10.62411/faith.2024-2.

[41] R. Suhendra *et al.*, "Cardiovascular Disease Prediction Using Gradient Boosting Classifier," *Infolitika J. Data Sci.*, vol. 1, no. 2, pp. 56–62, Dec. 2023, doi: 10.60084/ijds.v1i2.131.

[42] O. Jaiyeoba, E. Ogbuju, O. T. Yomi, and F. Oladipo, "Development of a Model to Classify Skin Diseases using Stacking Ensemble Machine Learning Techniques," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 22–38, May 2024, doi: 10.62411/jcta.10488.

[43] F. Mustofa, A. N. Safriandono, A. R. Muslikh, and D. R. I. M. Setiadi, "Dataset and Feature Analysis for Diabetes Mellitus Classification using Random Forest," *J. Comput. Theor. Appl.*, vol. 1, no. 1, pp. 41–48, Jan. 2023, doi: 10.33633/jcta.v1i1.9190.

[44] D. R. I. M. Setiadi, K. Nugroho, A. R. Muslikh, S. W. Iriananda, and A. A. Ojugo, "Integrating SMOTE-Tomek and Fusion Learning with XGBoost Meta-Learner for Robust Diabetes Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 23–38, May 2024, doi: 10.62411/faith.2024-11.

[45] T. R. Noviandy, G. M. Idroes, I. Hardi, M. Afjal, and S. Ray, "A Model-Agnostic Interpretability Approach to Predicting Customer Churn in the Telecommunications Industry," *Infolitika J. Data Sci.*, vol. 2, no. 1, pp. 34–44, May 2024, doi: 10.60084/ijds.v2i1.199.

[46] T. R. Noviandy, G. M. Idroes, M. Syukri, and R. Idroes, "Interpretable Machine Learning for Chronic Kidney Disease Diagnosis: A Gaussian Processes Approach," *Indones. J. Case Reports*, vol. 2, no. 1, pp. 24–32, Jun. 2024, doi: 10.60084/ijcr.v2i1.204.

[47] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, and G. Geleijnse, "Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival," *Sci. Rep.*, vol. 11, no. 1, p. 6968, Mar. 2021, doi: 10.1038/s41598-021-86327-7.

[48] T. R. Noviandy, G. M. Idroes, and I. Hardi, "Enhancing Loan Approval Decision-Making: An Interpretable Machine Learning Approach Using LightGBM for Digital Economy Development," *Malaysian J. Comput.*, vol. 9, no. 1, pp. 1734–1745, Apr. 2024, doi: 10.24191/mjoc.v9i1.25691.

[49] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges," in *Communications in Computer and Information Science*, Springer, Cham, 2020, pp. 417–431. doi: 10.1007/978-3-030-65965-3_28.

[50] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[51] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?,"" in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[52] G. Hooker, L. Mentch, and S. Zhou, "Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance," *Stat. Comput.*, vol. 31, no. 6, p. 82, Nov. 2021, doi: 10.1007/s11222-021-10057-z.