Optimizing Hepatitis C Virus Inhibitor Identification with LightGBM and Tree-structured Parzen Estimator Sampling

Teuku Rizky Noviandy
Master Program in Artificial Intelligence, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia
trizkynoviandy@gmail.com

Ghifari Maulana Idroes
Department of Nuclear Engineering and Engineering Physics, Universitas Gadjah Mada, Yogyakarta, Indonesia
ghifarimaulana145@gmail.com

Aga Maulana
Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh, Indonesia
agamaulana@usk.ac.id

Razief Perucha Fauzie Afidh
Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh, Indonesia
razief@usk.ac.id

Rinaldi Idroes
Department of Pharmacy, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh, Indonesia
rinaldi.idroes@usk.ac.id (corresponding author)

Received: 8 September 2024 | Revised: 25 October 2024 and 9 October 2024 | Accepted: 29 October 2024

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | DOI: https://doi.org/10.48084/etasr.8947

ABSTRACT

Identifying potent inhibitors against the Hepatitis C Virus (HCV) is crucial due to the continuous emergence of drug-resistant strains. Traditional drug discovery methods, including high-throughput screening, are often resource-intensive and time-consuming. Machine Learning (ML) approaches, particularly Quantitative Structure-Activity Relationship modeling, have been increasingly adopted to address this. This study utilized LightGBM, an efficient gradient-boosting framework, to predict the activity of potential HCV inhibitors. Additionally, the Tree-structured Parzen Estimator (TPE) was employed for hyperparameter optimization to enhance model performance. The optimized LightGBM-TPE model outperformed other ML models, including standard LightGBM, XGBoost, Random Forest, K-Nearest Neighbors, and Support Vector Machines, achieving an accuracy of 86.27%, a precision of 85.47%, a recall of 87.50%, a specificity of 85.03%, and an F1-score of 86.47%. Feature importance analysis identified critical molecular descriptors contributing to the model's predictive power. The results underscore the potential of advanced ML techniques and robust optimization methods to accelerate drug discovery, particularly for challenging targets such as HCV.

Keywords-classification; drug discovery; machine learning; QSAR modeling; supervised learning

I. INTRODUCTION

Hepatitis C Virus (HCV) is a major global health threat, affecting approximately 58 million people worldwide, with nearly 1.5 million new infections annually [1, 2]. Unlike other types of Hepatitis (such as Hepatitis A or B), Hepatitis C is unique in its ability to lead to chronic infection in most infected individuals, significantly increasing the risk of severe liver diseases, including cirrhosis and hepatocellular carcinoma [3]. Chronic HCV infections account for a substantial proportion of liver transplants and liver-related mortality globally [4]. Although other forms of hepatitis may have vaccines or a less frequent progression to chronic disease, the global burden of HCV, combined with the lack of a vaccine, makes it an urgent target for drug discovery efforts [5]. Despite the development of Direct-Acting Antivirals (DAAs), drug-resistant HCV strains continue to emerge, compromising treatment effectiveness [6]. This ongoing challenge requires the discovery and development of new inhibitors to maintain effective treatment options for patients.

Traditional High-Throughput Screening (HTS) has been a valuable method in drug discovery, but it is often resource-intensive, time-consuming, and costly [7]. Given the vast chemical space that needs to be explored to identify potential drug candidates, HTS presents significant limitations in efficiency. Consequently, researchers are increasingly seeking more efficient alternatives, with Machine Learning (ML) emerging as a promising solution to accelerate and enhance the drug discovery process [8, 9].

Quantitative Structure-Activity Relationship (QSAR) modeling is a widely used approach in drug discovery that has advanced significantly with the integration of ML. QSAR aims to establish correlations between chemical structures and their biological activities, and although traditional methods rely on statistical techniques, modern QSAR increasingly incorporates ML algorithms to enhance predictive power [10, 11]. The availability of large-scale biological and chemical datasets, along with advances in computational power, has propelled the adoption of ML in drug discovery. Early QSAR models were linear, but modern approaches now incorporate sophisticated, nonlinear ML techniques, greatly improving predictive capabilities. ML algorithms such as Random Forests (RF) [12], Support Vector Machines (SVM) [13], and deep neural networks [14] have proven effective, while recent works, such as [15, 16], using gradient-boosting algorithms have further improved accuracy and robustness. These innovations have enhanced the prediction of compound efficacy and accelerated the identification of promising drug candidates, highlighting the growing impact of ML-driven QSAR models.

The application of ML in QSAR modeling offers several advantages over conventional approaches. First, prioritizing compounds for experimental testing can significantly reduce the time and cost associated with early-stage drug discovery [17]. Second, ML-based QSAR models can handle complex, non-linear relationships between molecular structures and biological activities, potentially uncovering novel structure-activity patterns that traditional analytical methods might miss [18]. Finally, since these models are continuously refined with new data, their predictive power tends to improve over time, making them increasingly valuable tools in the drug discovery pipeline [19].

LightGBM, a gradient-boosting framework, has gained significant popularity due to its efficiency, scalability, and accuracy [20-22], particularly in domains such as drug discovery [23, 24], where its ability to handle large datasets and capture complex patterns in chemical-biological activity relationships has been demonstrated by studies such as [25, 26]. Compared to deep neural networks and other gradient-boosting frameworks, such as XGBoost [27] and CatBoost, which often require substantial computational resources for optimal performance, LightGBM provides a faster and more efficient approach [28], making it suitable for drug discovery tasks where rapid iteration is critical.

Despite its advantages, the performance of LightGBM models is highly dependent on the selection of optimal hyperparameters [29]. Traditional tuning methods, such as random and grid search, often struggle with computational efficiency and fail to explore the hyperparameter space effectively [30, 31], posing challenges in achieving optimal results. To address this, the Tree-structured Parzen Estimator (TPE), a sequential model-based optimization technique, has emerged as a robust alternative for hyperparameter optimization [32]. Introduced as an efficient method for modeling complex high-dimensional hyperparameter spaces [33], TPE has been shown to significantly improve the predictive accuracy and robustness of LightGBM models [34-35], making it especially valuable for identifying potent HCV inhibitors. Compared to other methods, such as Gaussian processes, TPE is better suited for complex, high-dimensional spaces due to its approach of separately modeling the distributions of good and poor configurations.

This study aims to develop a reliable and efficient QSAR model using LightGBM to predict the activity of potential HCV inhibitors based on their chemical structures. The key innovation is the use of LightGBM combined with the TPE for hyperparameter tuning, which improves the model's predictive accuracy and efficiency compared to traditional methods such as grid or random search. Using LightGBM's speed and accuracy, this study aims to enhance QSAR modeling in drug discovery. Additionally, the study focuses on optimizing the model's hyperparameters with the TPE method, which efficiently explores the hyperparameter space to further boost LightGBM's performance. This approach is expected to speed up the identification of effective HCV inhibitors and streamline the drug discovery process.

II. METHODS

A. Data Collection and Preprocessing

This study employed a dataset from [36], which focuses on bioactive compounds targeting the HCV NS5B protein, a critical enzyme in the RNA genome replication of the HCV. The dataset comprises 1,671 chemical compounds and includes the pIC_{50} values as the target variable. The pIC_{50} is the negative logarithm of the half-maximal inhibitory concentration, representing the potency of a compound in inhibiting a specific biological or biochemical function [37, 38].

As features for the proposed model, the molecular descriptors were calculated using AlvaDesc [39] within the Online Chemical Modelling Environment (OCHEM) [40], generating a total of 5,668 descriptors for each compound. To ensure the robustness of the model, multicollinearity was addressed by removing descriptors with a Pearson correlation coefficient greater than 0.95. This process resulted in a refined set of 2,468 descriptors. Subsequently, the data were normalized using a standard scaler [41]. Finally, the dataset was split into training and test sets, with 80% of the data allocated to training and the remaining 20% for testing [42]. Table I shows the distribution of active and inactive compounds in each subset [43]. The training set consists of 1336 compounds, with 671 active and 665 inactive compounds, while the testing set includes 335 compounds, of which 168 are active and 167 are inactive. This relatively balanced distribution ensures that the model is adequately trained and evaluated on both active and inactive compounds.

TABLE I. DISTRIBUTION OF ACTIVE AND INACTIVE COMPOUNDS IN TRAINING AND TESTING SETS

Subset	Active	Inactive	Total
Training Set	671	665	1336
Testing Set	168	167	335
Total			**1671**

B. *QSAR Modeling with LightGBM and Tree-structured Parzen Estimator (TSE)*

To model the relationship between molecular descriptors and pIC_{50} values, LightGBM was used, which is a gradient-boosting framework that has proven to be highly effective for a variety of ML tasks [22]. LightGBM is particularly well-suited to handle large-scale data and offers excellent performance and speed due to its histogram-based learning algorithm [44]. The LightGBM algorithm is designed to minimize a specified loss function by constructing an ensemble of decision trees, where each tree corrects the errors of the preceding ones [45]. The fundamental concept behind the LightGBM boosting mechanism is shown in:

$$y_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \qquad (1)$$

where y_i is the predicted output for the i^{th} sample, K is the number of boosting rounds, f_k represents each tree in the ensemble, x_i is the input feature vector, and $\mathcal{F}$ is the space of decision trees.

The TPE approach was employed for hyperparameter optimization to ensure that the LightGBM model is optimally tuned for the data. The TPE optimization process involves finding the set of hyperparameters that minimize the model prediction error on a validation set [46]. The TPE algorithm evaluates hyperparameters by modeling the probability of a good hyperparameter configuration. Compared to traditional hyperparameter tuning methods such as grid search and random search, TPE offers several advantages. Grid search exhaustively searches through all hyperparameter combinations, which can be computationally expensive and inefficient, especially for large parameter spaces. Random search, while faster, selects hyperparameter combinations randomly, potentially missing better configurations. In contrast, TPE focuses on exploring promising regions of the hyperparameter space by modeling the probability of improvement, making it more efficient. This results in faster convergence to optimal configurations and improved performance.

Table II shows the hyperparameters and their ranges used in the optimization process. The optimization of these hyperparameters using TPE ensures that the LightGBM model is fine-tuned to achieve the best possible predictive performance on the QSAR task, thus enhancing the accuracy and reliability of the pIC_{50} value predictions. This approach not only improves accuracy but also aids in the generalization of the model to new, unseen data, which is a critical aspect of successful QSAR modeling.

TABLE II. HYPERPARAMETERS FOR LIGHTGBM OPTIMIZATION WITH TPE

Hyperparameter	Range
num_leaves	20 to 150
max_depth	3 to 15
learning_rate	1×10^{-4} to 1×10^{-1} (log)
n_estimators	50 to 1000
min_child_samples	5 to 100
subsample	0.5 to 1.0
colsample_bytree	0.5 to 1.0
reg_alpha	1×10^{-8} to 10.0 (log)
reg_lambda	1×10^{-8} to 10.0 (log)

C. *Model Evaluation*

The evaluation of QSAR models in predicting the activity of potential HCV inhibitors involves multiple performance metrics, namely accuracy, precision, recall, specificity, and F1-score [47-49]. These metrics are integral to understanding the models' predictive capabilities from different perspectives, allowing for a nuanced comparison of their effectiveness. The choice of these metrics ensures that the evaluation comprehensively covers the correctness of predictions and the model's ability to generalize across unseen data. The equations for these metrics are shown below.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (2)$$

$$Precision = \frac{TP}{TP+FP} \qquad (3)$$

$$Recall = \frac{TP}{TP+FN} \qquad (4)$$

$$Specificity = \frac{TN}{TN+FP} \qquad (5)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (6)$$

where *TP* refers to the number of correctly predicted positive cases, *TN* is the number of correctly predicted negative cases, *FP* represents the number of negative cases incorrectly predicted as positive, and *FN* indicates the number of positive cases incorrectly predicted as negative.

To provide a comprehensive evaluation, the performance of LightGBM-TPE was compared to several other ML algorithms using their default hyperparameters: standard LightGBM, XGBoost, RF, K-Nearest Neighbors (KNN), and SVM. Each of these algorithms brings unique strengths to the task of QSAR modeling. Specifically, the comparison to LightGBM without TPE allows us to assess the impact of TPE on LightGBM's performance and determine how much of the observed performance gain is attributable to TPE.

III. RESULTS AND DISCUSSION

The hyperparameter optimization process using TPE resulted in a notable enhancement in the performance of the LightGBM model in predicting HCV inhibitor efficacy. Figure 1 shows the objective value (model accuracy) for 100 trials. The blue dots represent the objective value for each trial, while the red line indicates the best value achieved up to that point. The optimization stabilizes after approximately 11 trials, with the best objective value exceeding 0.86. This indicates that TPE identified an optimal hyperparameter configuration early in the process, leading to a robust and accurate QSAR model.

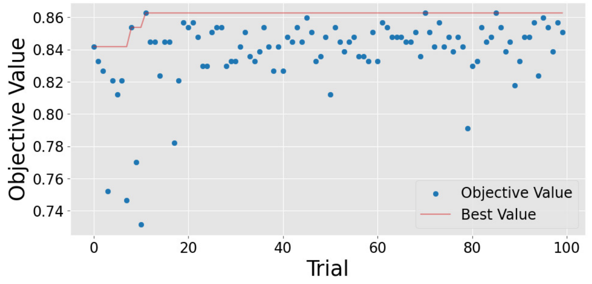

Fig. 1. Objective value optimization across 100 trials using LightGBM-TPE.

TABLE III. BEST HYPERPARAMETERS FOR LIGHTGBM-TPE

Hyperparameter	Value
num_leaves	146
max_depth	15
learning_rate	0.02
n_estimators	714
min_child_samples	64
subsample	0.62
colsample_bytree	0.52
reg_alpha	0.00
reg_lambda	0.004

Table III outlines the optimal hyperparameters determined for the LightGBM model using the TPE optimization approach in trial 11. These hyperparameters include a *maximum depth* of 15 and 146 leaves, which control the complexity of the trees, as well as a *learning rate* of 0.02, which regulates the step size during model training. Additionally, 714 estimators were used to control the number of boosting iterations, while the *subsample* and *colsample_bytree* parameters, set at 0.62 and 0.52 respectively, help prevent overfitting by introducing randomness in row and feature sampling. The regularization terms, *reg_alpha* and *reg_lambda*, were tuned to very low values, ensuring minimal penalization, which likely indicates that overfitting was not a significant concern.

Figure 2 illustrates the importance of various hyperparameters in optimizing the objective value for the LightGBM-TPE model. The learning rate emerges as the most critical factor, contributing 47% to the model's performance. *Colsample_bytree* and *n_estimators* follow with contributions of 22% and 14%, respectively. Other hyperparameters, such as *min_child_samples* and *subsample*, account for 7% each, while *num_leaves* has a minimal contribution of 1%. Hyperparameters like *max_depth*, *reg_lambda*, and *reg_alpha* contribute less than 1%, indicating their negligible impact on the model's performance. This analysis highlights the significant influence of the learning rate on the model's accuracy, making it a key focus in the hyperparameter tuning process. The prominence of *colsample_bytree* and *n_estimators* further underscores the importance of optimizing tree structure and ensemble size to achieve the best model performance.

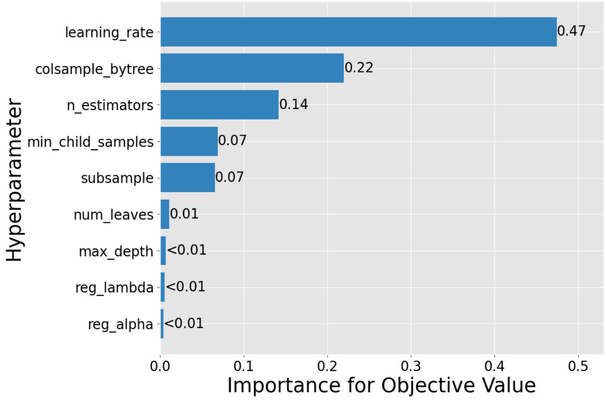

Fig. 2. Importance of hyperparameters for the objective value in the LightGBM-TPE model.

Table IV compares different ML models used to predict the efficacy of HCV inhibitors, focusing on key performance metrics. LightGBM-TPE demonstrates superior performance across all of these metrics. Specifically, the LightGBM-TPE model achieves an accuracy of 86.27%, precision of 85.47%, recall of 87.50%, specificity of 85.03%, and an F1-score of 86.47%. Other models such as standard LightGBM, XGBoost, RF, KNN, and SVM were also evaluated for comparison. XGBoost, a widely used gradient boosting algorithm, performs well with an accuracy of 82.09% and a balanced precision and recall, leading to an F1-score of 82.04%. However, it falls short of the LightGBM-TPE model, particularly in recall, indicating that it may miss more true positives in the prediction process.

TABLE IV. PERFORMANCE METRICS OF DIFFERENT MODELS IN PREDICTING HCV INHIBITOR EFFICACY.

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-score (%)
LightGBM-TPE	86.27	85.47	87.50	85.03	86.47
LightGBM	84.48	84.94	83.93	85.03	84.43
XGBoost	82.09	82.53	81.55	82.63	82.04
RF	83.28	81.46	86.31	80.24	83.82
KNN	82.99	81.01	86.31	79.64	83.57
SVM	81.79	80.23	84.52	79.04	82.32

RF performed slightly better than XGBoost in terms of recall (86.31%) and achieved an accuracy of 83.28%. However, its lower precision of 81.46% resulted in an F1-score of 83.82%. KNN followed closely with an accuracy of 82.99% and an F1-score of 83.57%, indicating comparable performance to RF but still not reaching the level of LightGBM-TPE. SVM shows the lowest performance among the models, with an accuracy of 81.79% and an F1-score of 82.32%. The lower precision of 80.23% suggests that SVM is more prone to false positives than the other models.

LightGBM-TPE not only outperformed the other models but also exceeded previous studies that utilized stacked classifiers to predict HCV inhibitors, which achieved a top accuracy of 85.07% [50]. By achieving an accuracy of 86.27%, LightGBM-TPE demonstrates a clear improvement over stacked classifiers, further highlighting the advantages of TPE optimization. This enhanced accuracy, along with superior precision, recall, and F1-score, underscores the potential of LightGBM-TPE as a more reliable and effective approach to identifying active compounds in drug discovery.

Table V provides a detailed comparison of actual versus predicted outcomes for the various models classifying compounds as active or inactive HCV inhibitors. The LightGBM-TPE model demonstrated superior performance, correctly identifying 147 active and 142 inactive compounds, with only 21 false negatives and 25 false positives. This balance between true positive and true negative predictions highlights the model's effectiveness in sensitivity (recall) and specificity, which is critical in reducing both classification errors. In comparison, the standard LightGBM model correctly identified 141 active and 142 inactive compounds, with 27 false negatives and 25 false positives, indicating a slightly lower accuracy than LightGBM-TPE, especially in identifying active compounds. XGBoost also fell short, correctly predicting 137 active and 138 inactive compounds, but with higher false negatives (31) and false positives (29), showing less accuracy compared to both versions of LightGBM. RF and KNN correctly classified 145 active compounds, but RF showed 23 false negatives and 33 false positives while KNN had 23 false negatives and 34 false positives. Both models underperformed compared to LightGBM-TPE, particularly in distinguishing inactive compounds. Lastly, SVM performed the weakest, with 142 correct predictions for active compounds and 132 for inactive compounds, along with the highest number of false negatives (26) and false positives (35), making it the least accurate.

TABLE V. CONFUSION MATRIX OF DIFFERENT MODELS IN PREDICTING HCV INHIBITOR EFFICACY

Model	**Actual**	**Predicted**	
		Active	**Inactive**
LightGBM-TPE	Active	147	21
	Inactive	25	142
LightGBM	Active	141	27
	Inactive	25	142
XGBoost	Active	137	31
	Inactive	29	138
RF	Active	145	23
	Inactive	33	134
KNN	Active	145	23
	Inactive	34	133
SVM	Active	142	26
	Inactive	35	132

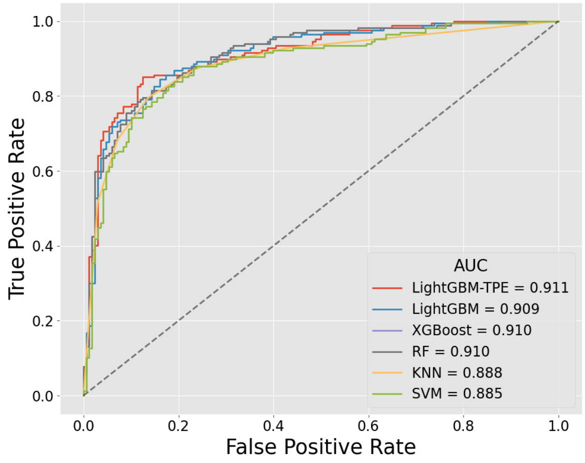

Fig. 3. ROC curves across various ML models.

Figure 3 illustrates the Receiver Operating Characteristic (ROC) curves for the five models, with the Area Under the Curve (AUC) values displayed in the legend. The ROC curve plots the True Positive Rate (sensitivity) against the False Positive Rate, visually representing the trade-offs between sensitivity and specificity for each model. The LightGBM-TPE model achieved the highest AUC of 0.911, indicating its superior ability to distinguish between active and inactive compounds. It was closely followed by XGBoost and RF, with AUC values of 0.910, and standard LightGBM with AUC values of 0.909. demonstrating comparable performance. KNN and SVM, with AUC values of 0.888 and 0.885, respectively, show slightly lower performance, which aligns with the previous evaluation metrics. The ROC curves for LightGBM-TPE, XGBoost, and RF closely overlap, reflecting their strong and similar performance in classification tasks. In contrast, the ROC curves for KNN and SVM are slightly lower, particularly in the early stages of the curve, indicating that these models are more prone to false positives than LightGBM-TPE, XGBoost, and RF. This further validates the choice of LightGBM-TPE as the most effective model to predict HCV inhibitor efficacy, combining high sensitivity with a low rate of false positives.

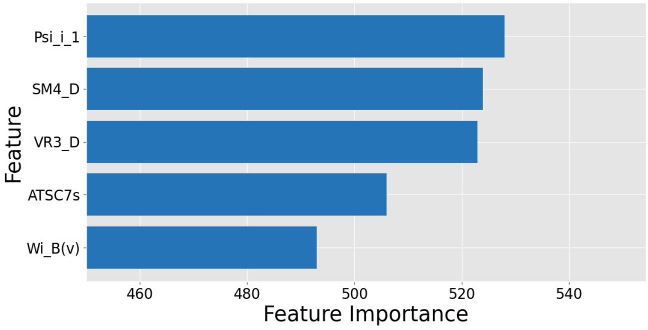

Fig. 4. Feature importance rankings for the top five features identified by the LightGBM-TPE model.

Figure 4 shows the five most important features identified by the LightGBM-TPE model in predicting the efficacy of HCV inhibitors. Feature importance was calculated using LightGBM's split-based method, which works by counting how often a feature is used to split the data across all trees in the model. Features that are used more frequently to create splits are assigned higher importance scores, as they contribute more to reducing the model's overall error. This approach provides a clear interpretation of feature relevance by directly linking it to the model's decision-making process during training.

Among these features, Psi_i_1 stands out as the most influential, followed closely by SM4_D and VR3_D. These features are likely critical molecular descriptors that play a key role in determining the biological activity of the compounds. ATSC7s and Wi_B(v) also contribute significantly, though slightly less than the top three. This analysis of feature importance offers valuable insights into which molecular properties are most closely associated with the activity of HCV inhibitors. Understanding these key features can guide further research, focusing on optimizing these specific aspects of chemical compounds to enhance their efficacy as potential drug candidates. The prominence of these features in the model's decision-making process underscores their relevance in the biological mechanisms targeted by the inhibitors.

The integration of the LightGBM model with TPE demonstrated significant improvements in the prediction accuracy of HCV inhibitors. Beyond the advancement in ML techniques, this study highlights the critical role of QSAR modeling in enhancing the drug discovery process. QSAR modeling, by establishing correlations between chemical structures and their biological activities, is fundamental in identifying structure-activity relationship patterns that drive the efficacy of drug candidates. In this context, the combination of LightGBM and QSAR allows efficient identification of potential inhibitors by leveraging molecular descriptors that capture key structural and chemical features relevant to the biological activity of the compounds.

The results highlight how advanced ML techniques, when applied to QSAR modeling, can optimize the identification of potent drug candidates. The LightGBM-TPE model consistently outperformed other ML models, such as standard LightGBM, XGBoost, RF, KNN, and SVM, in key performance metrics, including accuracy, precision, recall, and F1-score. This suggests that integrating sophisticated ML algorithms with QSAR can significantly accelerate the drug discovery process, enhancing the prediction of active compounds while reducing the time and resources required for early-stage screening. The importance of hyperparameter tuning through TPE further optimized model performance, underscoring the critical role of robust optimization techniques in predictive QSAR modeling.

A central aspect of this study is the contribution of QSAR in uncovering structure-activity relationship patterns, which are instrumental in predicting the inhibitory activity of compounds. The molecular descriptors used in QSAR capture various chemical, physical, and topological properties of the compounds, which play a pivotal role in determining their biological activity. For instance, in the case of HCV inhibitors, molecular descriptors such as electronic, hydrophobic, and steric properties may significantly affect how compounds interact with the target protein, ultimately influencing their inhibitory potency. By integrating thousands of molecular descriptors into the LightGBM model, the QSAR framework allows the identification of critical features that drive inhibitory activity against HCV.

This study also demonstrates how molecular descriptors enable the model to capture nonlinear relationships between chemical structures and biological activities, which may be difficult to detect using traditional linear QSAR methods. The use of LightGBM allows for the identification of complex interactions among multiple descriptors, enhancing the model's predictive power and enabling it to uncover novel structure-activity relationship patterns. This ability to model complex relationships is particularly valuable in drug discovery, where subtle variations in molecular structure can lead to significant differences in biological activity.

Despite promising results, this study has some limitations. First, the dataset is small and specific to HCV inhibitors, which may limit the application of the findings to other drug discovery areas. Expanding the dataset to include more diverse compounds and targets could make the model more robust. Second, although the LightGBM-TPE model showed good performance, its complexity might make it harder for researchers unfamiliar with advanced ML techniques to interpret. Using simpler models or interpretive methods could clarify how certain molecular features affect predictions. Finally, this study focuses on binary classification. Including multiclass or regression approaches could provide a fuller understanding of how chemical structures influence activity across a spectrum. Future research could explore alternative ML algorithms or hybrid models that combine the strengths of multiple approaches, further improving predictive accuracy and robustness. Additionally, further research into the role of specific molecular descriptors in determining biological activity could provide deeper insights into the structure-activity relationships that drive drug efficacy. This, in turn, could guide the design of more effective inhibitors by targeting the key molecular features that contribute to their potency.

IV. CONCLUSION

The application of LightGBM-TPE in this study demonstrated a significant advancement in the predictive modeling of HCV inhibitors, achieving outstanding performance metrics, with an accuracy of 86.27%, precision of 85.47%, recall of 87.50%, specificity of 85.03%, and F1-score of 86.47%, surpassing other models such as standard LightGBM, XGBoost, RF, KNN, and SVM. The novelty of this study lies in the use of TPE for optimizing LightGBM, which has not been extensively explored in QSAR modeling for HCV inhibitors. Compared to traditional grid and random search methods, TPE allows faster and more accurate hyperparameter tuning, resulting in enhanced model performance. The findings demonstrate that integrating advanced optimization methods with QSAR modeling can significantly improve the identification of potent drug candidates, reducing the time and computational resources required for early-stage drug discovery. This study contributes to the growing body of literature by offering a more efficient and accurate approach to predictive modeling in drug discovery, with potential applications beyond HCV inhibitors to other disease targets.

ACKNOWLEDGMENT

This research was funded by Universitas Syiah Kuala, Kementerian Pendidikan, Kebudayaan, Riset dan Teknologi through the "Penelitian Profesor" scheme, grant no. 18/UN11.2.1/PT.01.03/PNBP/2023.

REFERENCES

[1] P. Axley, Z. Ahmed, S. Ravi, and A. K. Singal, "Hepatitis C Virus and Hepatocellular Carcinoma: A Narrative Review," *Journal of Clinical and Translational Hepatology*, vol. 6, no. 1, Dec. 2017, Art. no. 79, https://doi.org/10.14218/JCTH.2017.00067.

[2] T. Stroffolini and G. Stroffolini, "Prevalence and Modes of Transmission of Hepatitis C Virus Infection: A Historical Worldwide Review," *Viruses*, vol. 16, no. 7, Jul. 2024, Art. no. 1115, https://doi.org/10.3390/v16071115.

[3] F. Fiehn, C. Beisel, and M. Binder, "Hepatitis C virus and hepatocellular carcinoma: carcinogenesis in the era of direct-acting antivirals," *Current Opinion in Virology*, vol. 67, Aug. 2024, Art. no. 101423, https://doi.org/10.1016/j.coviro.2024.101423.

[4] L. Gvinjilia *et al.*, "Impact of Hepatitis C Virus Infection and Treatment on Mortality in the Country of Georgia, 2015–2020," *Clinical Infectious Diseases*, vol. 77, no. 3, pp. 405–413, Aug. 2023, https://doi.org/10.1093/cid/ciad182.

[5] A. L. Cox *et al.*, "Progress towards elimination goals for viral hepatitis," *Nature Reviews Gastroenterology & Hepatology*, vol. 17, no. 9, pp. 533–542, Sep. 2020, https://doi.org/10.1038/s41575-020-0332-6.

[6] M. Bhatia and E. Gupta, "Emerging resistance to directly-acting antiviral therapy in treatment of chronic Hepatitis C infection—A brief review of literature," *Journal of Family Medicine and Primary Care*, vol. 9, no. 2, Feb. 2020, Art. no. 531, https://doi.org/10.4103/jfmpc.jfmpc_943_19.

[7] S. Singh, H. Gupta, P. Sharma, and S. Sahi, "Advances in Artificial Intelligence (AI)-assisted approaches in drug screening," *Artificial Intelligence Chemistry*, vol. 2, no. 1, Jun. 2024, Art. no. 100039, https://doi.org/10.1016/j.aichem.2023.100039.

[8] M. Elbadawi, S. Gaisford, and A. W. Basit, "Advanced machine-learning techniques in drug discovery," *Drug Discovery Today*, vol. 26, no. 3, pp. 769–777, Mar. 2021, https://doi.org/10.1016/j.drudis.2020.12.003.

[9] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, "Artificial intelligence to deep learning: machine intelligence approach for drug discovery," *Molecular Diversity*, vol. 25, no. 3, pp. 1315–1360, Aug. 2021, https://doi.org/10.1007/s11030-021-10217-3.

[10] T. R. Noviandy *et al.*, "Ensemble Machine Learning Approach for Quantitative Structure Activity Relationship Based Drug Discovery: A Review," *Infolitika Journal of Data Science*, vol. 1, no. 1, pp. 32–41, Sep. 2023, https://doi.org/10.60084/ijds.v1i1.91.

[11] T. R. Noviandy, G. M. Idroes, T. E. Tallei, D. Handayani, and R. Idroes, "QSAR Modeling for Predicting Beta-Secretase 1 Inhibitory Activity in Alzheimer's Disease with Support Vector Regression," *Malacca Pharmaceutics*, vol. 2, no. 2, pp. 79–85, Sep. 2024, https://doi.org/10.60084/mp.v2i2.226.

[12] 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, Jan. 2020, Art. no. 1317, https://doi.org/10.3390/molecules25061317.

[13] M. Fajar Rizqi, R. Rendian Septiawan, and I. Kurniawan, "Implementation of Simulated Annealing-Support Vector Machine on QSAR Study of Indenopyrazole Derivative as Anti-Cancer Agent," in *2021 9th International Conference on Information and Communication Technology (ICoICT)*, Yogyakarta, Indonesia, Aug. 2021, pp. 662–668, https://doi.org/10.1109/ICoICT52021.2021.9527416.

[14] C. Gui, Y. Li, and T. Peng, "Development of predictive QSAR models for the substrates/inhibitors of OATP1B1 by deep neural networks," *Toxicology Letters*, vol. 376, pp. 20–25, Mar. 2023, https://doi.org/10.1016/j.toxlet.2023.01.006.

[15] A. de F. Cobre *et al.*, "Identifying 124 new *anti*-HIV drug candidates in a 37 billion-compound database: An integrated approach of machine learning (QSAR), molecular docking, and molecular dynamics simulation," *Chemometrics and Intelligent Laboratory Systems*, vol. 250, Jul. 2024, Art. no. 105145, https://doi.org/10.1016/j.chemolab.2024.105145.

[16] Z. Zhao, J. Yang, H. Ji, Z. Liu, T. Sun, and T. NI, "QSAR Model based Gradient Boosting Regression of N-Arylsulfonyl-Indole-2-Carboxamide Derivatives as Inhibitors for Fructose-1,6-bisphosphatase," *Letters in Drug Design & Discovery*, vol. 21, no. 7, pp. 1274–1286, Jun. 2024, https://doi.org/10.2174/1570180820666230726145659.

[17] 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," *Journal of Soft Computing and Data Mining*, vol. 5, no. 1, pp. 46–56, Jun. 2024.

[18] A. Karampuri and S. Perugu, "A breast cancer-specific combinational QSAR model development using machine learning and deep learning approaches," *Frontiers in Bioinformatics*, vol. 3, Jan. 2024, https://doi.org/10.3389/fbinf.2023.1328262.

[19] H. Ding, F. Xing, L. Zou, and L. Zhao, "QSAR analysis of VEGFR-2 inhibitors based on machine learning, Topomer CoMFA and molecule docking," *BMC Chemistry*, vol. 18, no. 1, Mar. 2024, Art. no. 59, https://doi.org/10.1186/s13065-024-01165-8.

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

[21] C. Chen, Q. Zhang, Q. Ma, and B. Yu, "LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion," *Chemometrics and Intelligent Laboratory Systems*, vol. 191, pp. 54–64, Aug. 2019, https://doi.org/10.1016/j.chemolab.2019.06.003.

[22] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[23] C. Chen and H. Seo, "Prediction of rock mass class ahead of TBM excavation face by ML and DL algorithms with Bayesian TPE optimization and SHAP feature analysis," *Acta Geotechnica*, vol. 18, no. 7, pp. 3825–3848, Jul. 2023, https://doi.org/10.1007/s11440-022-01779-z.

[24] F. Hou, Z. Cheng, L. Kang, and W. Zheng, "Prediction of Gestational Diabetes Based on LightGBM," in *Proceedings of the 2020 Conference on Artificial Intelligence and Healthcare*, Taiyuan, China, Oct. 2020, pp. 161–165, https://doi.org/10.1145/3433996.3434025.

[25] M. Jalal, M. Kamal, and A. Zafar, "ChemCarcinoPred: Carcinogenicity Prediction of Small Drug-Like Molecules Using LightGBM and Molecular Fingerprints," *Biophysical Reviews and Letters*, pp. 1–16, Aug. 2023, https://doi.org/10.1142/S1793048023410035.

[26] M. Stawiski, P. Meier, R. Dornberger, and T. Hanne, "Using the Light Gradient Boosting Machine for Prediction in QSAR Models," in *Proceedings of International Joint Conference on Advances in Computational Intelligence*, 2023, pp. 99–111, https://doi.org/10.1007/978-981-99-1435-7_10.

[27] 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," *Journal of Future Artificial Intelligence and Technologies*, vol. 1, no. 1, pp. 51–63, Jun. 2024, https://doi.org/10.62411/faith.2024-12.

[28] J. Zhang, D. Mucs, U. Norinder, and F. Svensson, "LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity–Application to the Tox21 and Mutagenicity Data Sets," *Journal of Chemical Information and Modeling*, vol. 59, no. 10, pp. 4150–4158, Oct. 2019, https://doi.org/10.1021/acs.jcim.9b00633.

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

[30] L. Liao, H. Li, W. Shang, and L. Ma, "An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks," *ACM Transactions on Software Engineering and Methodology*, vol. 31, no. 3, pp. 1–40, Jul. 2022, https://doi.org/10.1145/3506695.

[31] M. Mwita, J. Mbelwa, J. Agbinya, and A. E. Sam, "The Effect of Hyperparameter Optimization on the Estimation of Performance Metrics in Network Traffic Prediction using the Gradient Boosting Machine Model," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10714–10720, Jun. 2023, https://doi.org/10.48084/etasr.5548.

[32] M. Liang *et al.*, "Improving Genomic Prediction with Machine Learning Incorporating TPE for Hyperparameters Optimization," *Biology*, vol. 11, no. 11, Nov. 2022, Art. no. 1647, https://doi.org/10.3390/biology 11111647.

[33] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for Hyper-Parameter Optimization," in *Advances in Neural Information Processing Systems*, 2011, vol. 24.

[34] M. Wang, J. Zhang, H. Li, B. Zhang, and Z. Yang, "Identification of mine water source based on TPE-LightGBM," *Scientific Reports*, vol. 14, no. 1, May 2024, Art. no. 12539, https://doi.org/10.1038/s41598-024-62413-4.

[35] X. Xiong *et al.*, "Application of LightGBM hybrid model based on TPE algorithm optimization in sleep apnea detection," *Frontiers in Neuroscience*, vol. 18, Feb. 2024, https://doi.org/10.3389/fnins.2024. 1324933.

[36] S. Kamboj, A. Rajput, A. Rastogi, A. Thakur, and M. Kumar, "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, pp. 3422–3438, Jan. 2022, https://doi.org/10.1016/j.csbj.2022.06.060.

[37] A. Thakur, A. Kumar, V. Sharma, and V. Mehta, "PIC50: An open source tool for interconversion of PIC50 values and IC50 for efficient data representation and analysis." bioRxiv, Art. no. 2022.10.15.512366, Oct. 18, 2022, https://doi.org/10.1101/2022.10.15.512366.

[38] T. R. Noviandy, G. M. Idroes, F. Mohd Fauzi, and R. Idroes, "Application of Ensemble Machine Learning Methods for QSAR Classification of Leukotriene A4 Hydrolase Inhibitors in Drug Discovery," *Malacca Pharmaceutics*, vol. 2, no. 2, pp. 68–78, Sep. 2024, https://doi.org/10.60084/mp.v2i2.217.

[39] A. Mauri, "alvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints," in *Ecotoxicological QSARs*, K. Roy, Ed. New York, NY, USA: Springer US, 2020, pp. 801–820.

[40] I. Sushko *et al.*, "Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information," *Journal of Computer-Aided Molecular Design*, vol. 25, no. 6, pp. 533–554, Jun. 2011, https://doi.org/10.1007/s10822-011-9440-2.

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

[42] M. Agustia *et al.*, "Application of Fuzzy Support Vector Regression to Predict the Kovats Retention Indices of Flavors and Fragrances," in *2022 International Conference on Electrical Engineering and Informatics (ICELTICs)*, Banda Aceh, Indonesia, Sep. 2022, pp. 13–18, https://doi.org/10.1109/ICELTICs56128.2022.9932124.

[43] 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," *Journal of Future Artificial Intelligence and Technologies*, vol. 1, no. 1, pp. 39–50, Jun. 2024, https://doi.org/10.62411/faith.2024-2.

[44] 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 Journal of Data Science*, vol. 2, no. 1, pp. 34–44, May 2024, https://doi.org/10.60084/ijds.v2i1.199.

[45] A. Maulana *et al.*, "Performance Analysis and Feature Extraction for Classifying the Severity of Atopic Dermatitis Diseases," in *2023 2nd International Conference on Computer System, Information Technology, and Electrical Engineering (COSITE)*, Banda Aceh, Indonesia, Aug. 2023, pp. 226–231, https://doi.org/10.1109/COSITE60233.2023. 10249760.

[46] T. T. Khoei, S. Ismail, and N. Kaabouch, "Boosting-based Models with Tree-structured Parzen Estimator Optimization to Detect Intrusion Attacks on Smart Grid," in *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, USA, Dec. 2021, pp. 0165–0170, https://doi.org/10.1109/UEMCON53757.2021.9666607.

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

[48] D. J. I. Supriatna, H. Saputra, and K. Hasan, "Enhancing the Red Wine Quality Classification Using Ensemble Voting Classifiers," *Infolitika Journal of Data Science*, vol. 1, no. 2, pp. 42–47, Oct. 2023, https://doi.org/10.60084/ijds.v1i2.95.

[49] K. T. Nguyen, T. N. Tran, and H. T. Nguyen, "Research on the Influence of Hyperparameters on the LightGBM Model in Load Forecasting," *Engineering, Technology & Applied Science Research*, vol. 14, no. 5, pp. 17005–17010, Oct. 2024, https://doi.org/10.48084/etasr.8266.

[50] 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)*, Banda Aceh, Indonesia, Aug. 2023, pp. 220–225, https://doi.org/10.1109/COSITE60233.2023.10250039.