Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 6, No. 1, January 2024, pp: 1-10; eISSN: 2656-8632**

RESEARCH ARTICLE OPEN ACCESS

Manuscript received November 1, 2023; revised November 14, 2023; accepted November 15, 2023; date of publication January 20, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeemi.v6i1.338

How to cite: Daniel Febrian Sengkey, Angelina Stevany Regina Masengi, Regression Algorithms in Predicting the SARS-CoV-2 Replicase Polyprotein 1ab Inhibitor: A Comparative Study, vol. 6, no. 1, pp. 1-10, January 2024.

Regression Algorithms in Predicting the SARS-CoV-2 Replicase Polyprotein 1ab Inhibitor: A Comparative Study

Daniel Febrian Sengkey[1], Angelina Stevany Regina Masengi[2]
[1] Department of Electrical Engineering, Faculty of Engineering, Universitas Sam Ratulangi, Jl. Kampus Unsrat, Bahu, Manado 95115, North Sulawesi, INDONESIA
[2] Department of Pharmacology and Therapy, Faculty of Medicine, Universitas Sam Ratulangi, Jl. Kampus Unsrat, Bahu, Manado 95115, North Sulawesi, INDONESIA

Corresponding author: Daniel Febrian Sengkey (e-mail: danielsengkey@unsrat.ac.id).

ABSTRACT Due to its extensive steps and trials, drug discovery is a long and expensive process. In the last decade, as also hard pressed by the COVID-19 pandemic, the screening process could be assisted with the advancement in computational technology including the application of Machine Learning. The classification task in Machine Learning has become one of the major approaches for drug discovery. Unfortunately, this practice uses discretized labels that might lead to the loss of quantitative properties that could be meaningful. Therefore, in this paper, we aim to compare various Machine Learning regression algorithms in predicting inhibitory bioactivity, specifically the IC_{50} value, with the SARS-CoV-2 Replicase Polyprotein 1ab as the target. With 1,138 non-duplicated data downloaded from the ChEMBL database that was engineered into four dataset variances, 42 regression algorithms were utilized for the prediction. We found that there are computational challenges to the use of regression algorithms in predicting bioactivity, for only a handful and a specific dataset variance that returned valid performance parameters upon testing. The three that yielded the highest counts of valid performance parameters are the Histogram Gradient Boosting Regressor (HGBR), Light Gradient Boosting Machine Regressor (LGBR), and Random Forest Regression (RFR). Further statistical analyses show that there is no significant difference between these three algorithms, except for the time taken for training and testing the model, where the LGBR excels. Therefore, these three algorithms should be primarily considered for the study with the same nature.

INDEX TERMS SARS-CoV-2 replicase polyprotein 1ab, drug discovery, regression algorithms, IC_{50} prediction.

I. INTRODUCTION

It is well known that drug discovery is a both lengthy and costly process [8]. The steps that cover pre-clinical, including several trials are found to be similar in several developed countries and may take 11.5 years until a market introduction [9], [10]. However, in late 2019 a new form of coronavirus started infecting humans and even led to a global catastrophe in early 2020 [11]. The catastrophic circumstances brought a dire need for the accelerated discovery of novel therapeutic options such as new healthcare technology [12] and Computer Aided Drug Discovery (CADD) [13].

In silico drug discovery, where computational powers are utilized, is a solution that has been extensively studied in recent years, especially in the pre-clinical step [10], [13]–[16]. The methods include molecular dynamic simulation (MDS) [14] and machine learning [14]–[16]. It has been used for discovering new compounds or repurposing known drugs, and until 2021, there were more than 70 approved drugs that were utilized in silico methods in its pre-clinical stage [14].

COVID-19 caused by the SARS-CoV-2 virus is one of the diseases that received global attention and where the CADD is actively researched [17]. The drug candidate for SARS-CoV-2 may target human cells or, at the other end, target the virus itself [18]. In this case, the 3-chymotrypsin-like protease (3CLpro) or the Main Protease (Mpro), which is one of the key proteases of the SARS-CoV-2 that is important in replication is an enticing target due to the inexistence of its human homolog [19], [20]. This is the common scenario as reported

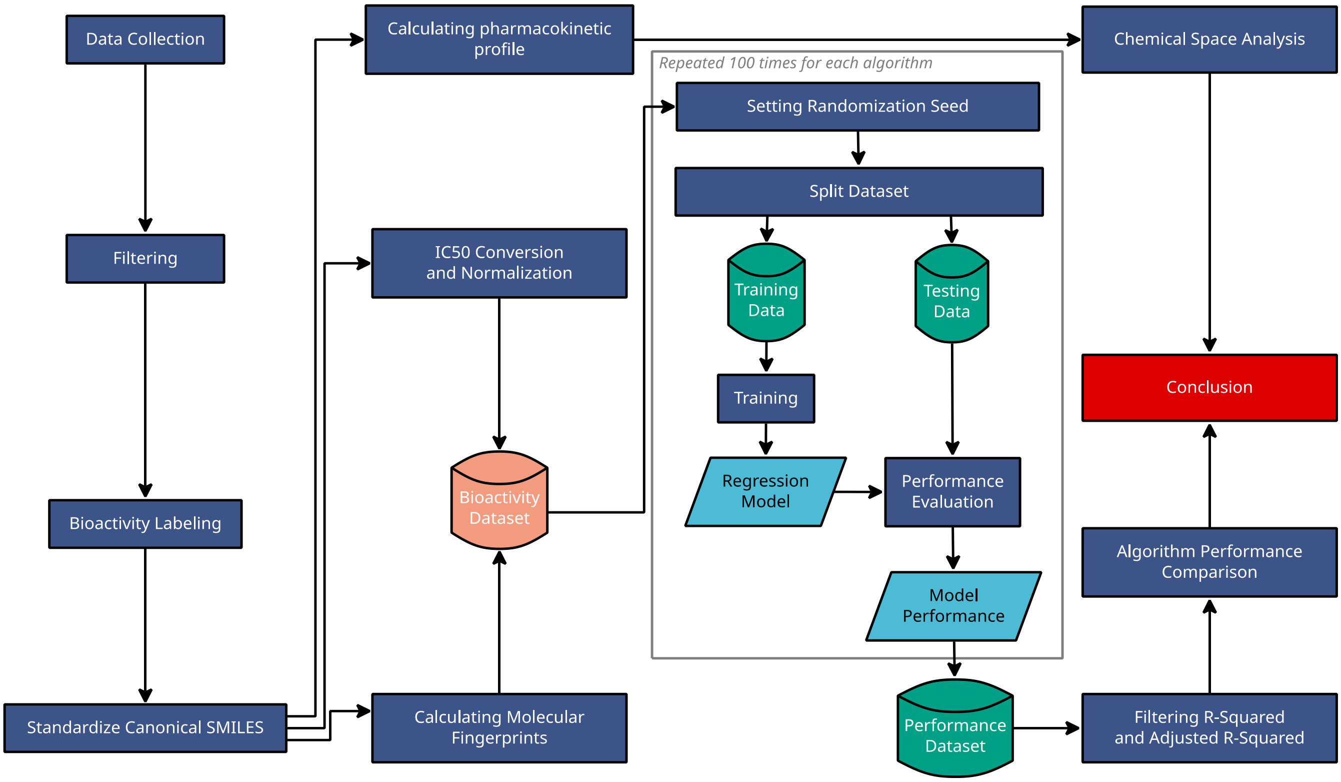

Data Collection
Filtering
Bioactivity Labeling
Standardize Canonical SMILES
Calculating pharmacokinetic profile
IC50 Conversion and Normalization
Bioactivity Dataset
Calculating Molecular Fingerprints
Repeated 100 times for each algorithm
Setting Randomization Seed
Split Dataset
Training Data
Testing Data
Training
Regression Model
Performance Evaluation
Model Performance
Performance Dataset
Chemical Space Analysis
Conclusion
Algorithm Performance Comparison
Filtering R-Squared and Adjusted R-Squared

FIGURE 1. Research Course

in various studies [20]–[25]. Another scenario targets the transmembrane protease serine 2 (TMPRSS2) at the host, rather than targeting the virus protein due to the highly mutative nature of the virus [26].

Machine Learning is a progressing field study concerning how computers can construct knowledge from a set of data (experience) [27]. It has been used to predict interactions between drugs and proteins, discover efficacy, and confirm the safety biomarkers [28]. Sulistiawan et al. [29] used a Deep Semi-Supervised Model to predict the Drug-Target Interaction (DTI) of the antiviral drug candidate of SARS-CoV-2. Machine Learning has also been used to discover the potential of Indonesian herbal compounds, as well as Traditional Chinese Medicine (TCM) as antiviral agents for SARS-CoV-2 [5], [15], [23]. In a broader sense, Machine Learning also was used to predict drug synergy in cancer cells [30], as well as empowering web servers that can be used to predict inhibitory activities [2]–[4].

In terms of tasks, classification is the common one, applied in drug discovery. Drugs with known interaction will be labeled as 1 and those with no interaction will be labeled as 0 [2]–[5], [29]–[32]. The label itself might have come from the DTI database or based on the half maximal inhibitory concentration value (IC_{50})[33], where the concentrations that are less than 1000 nM, between 1000 and 10000 nM, and greater than 10000 nM are labeled as active, intermediate, and inactive, respectively. In many cases, the intermediate ones are omitted.

Regardless of its popularity in drug discovery, the use of discretized labels based on IC_{50} reduces the data resolution, as well as complicates comparisons and evaluations with other numeric parameters. Grouping of numeric data is discouraged in epidemiology studies [34]. Therefore, in this study, we use an uncommon approach in drug discovery, where we apply prediction instead of classification. To gain better insights, we apply various regression algorithms with the IC_{50} value used as a label. The purpose is to evaluate the possibility of using regression in place of classification in Machine Learning-based drug discovery. The motivation is to preserve the quantitative properties in the label, which are commonly lost when the label is discretized into groups in classification studies. The availability of the predicted inhibitory bioactivity in its original quantitative form should provide a better chance of comparison with other drug discovery approaches in the future. The SARS-CoV-2 Replicase Polyprotein 1ab is used as our target protein since we are still moving on from the pandemic to epidemic status. The remainder of this article is organized as follows: in Section II we present the workflow as well as steps involved in this research, followed by the Results and Discussion in Section III. In Section IV the paper is concluded and some potential issues to be explored are presented.

II. METHODS

The method is similar to common data science methodology with some addition of the Chemical Space Analysis, which is common in drug discovery, and statistical comparison of the

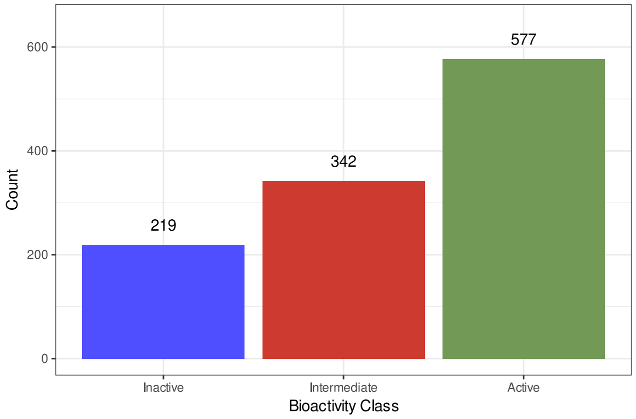


FIGURE 2. Compounds frequency in each bioactivity class.

algorithms [2], [3], [24], [35]. The whole process, as illustrated in FIGURE 1, comprised three main parts: data preparation, Quantitative Structure-Activity Relationship (QSAR) modeling and evaluation, and statistical analysis.

A. DATA PREPARATION

This part includes data collection and preprocessing. For this part, we used a free Google Colab service. The bioactivity dataset is queried from ChEMBL by using its web service and the provided Python *Application Programming Interface* (API) [1]. The ChEMBL ID for the target is CHEMBL4523582, queried on June 25th, 2023. At the moment, there were 1,220 inhibitory bioactivity data of the SARS-CoV-2 Replicase Polyprotein 1ab stored in the ChEMBL database. The queried data includes fields such as the compounds in Simplified Molecular-input-line-entry system (SMILES) notation and the IC_{50} value towards the target of each compound. Later, the duplicated compounds are filtered, resulting in 1,138 non-duplicated inhibitory bioactivity data. The next step is labeling each compound bioactivity as “active”, “intermediate”, or “inactive”, based on the respective IC_{50}. The values less than 1000 nM are labeled as active, while those greater than 10000 nM are labeled as inactive, and those in between are labeled as intermediate. The frequency of each bioactivity class is shown in FIGURE 2.

The SMILES representation of each compound is standardized using string processing. The standardized SMILES is then used for calculating the pharmacokinetic profile of each compound by using the Python rdkit library version 3.1 [36], [37]. The calculated profile includes the number of hydrogen bond donors (NumHDonors), the number of hydrogen bond acceptors (NumHAcceptors), molecular weight (MW), and the Ghose-Crippen-Viswanadhan octanol-water partition coefficient (LogP). Standardized SMILES are also used to calculate Pubchem molecular fingerprints, by using the function from the Padelpy library [38]. Then, the IC_{50} values are converted to pIC_{50}, and along with the standardized SMILES, and PubChem fingerprints, form the bioactivity dataset.

B. QSAR MODELING AND EVALUATION

The bioactivity from the previous step was then used as the dataset for training and testing the regression algorithms. The dataset was split with an 80:20 proportion for the training and testing data, respectively using the Scikit-learn Library [39]. Then, we used the lazypredict library for batch training and evaluation of the regression models [40]. In total, 42 regression algorithms were tested. The dataset splitting, model training, and evaluation were repeated 100 times, with the iteration number also serving as the randomization seed. Four dataset variances were used:

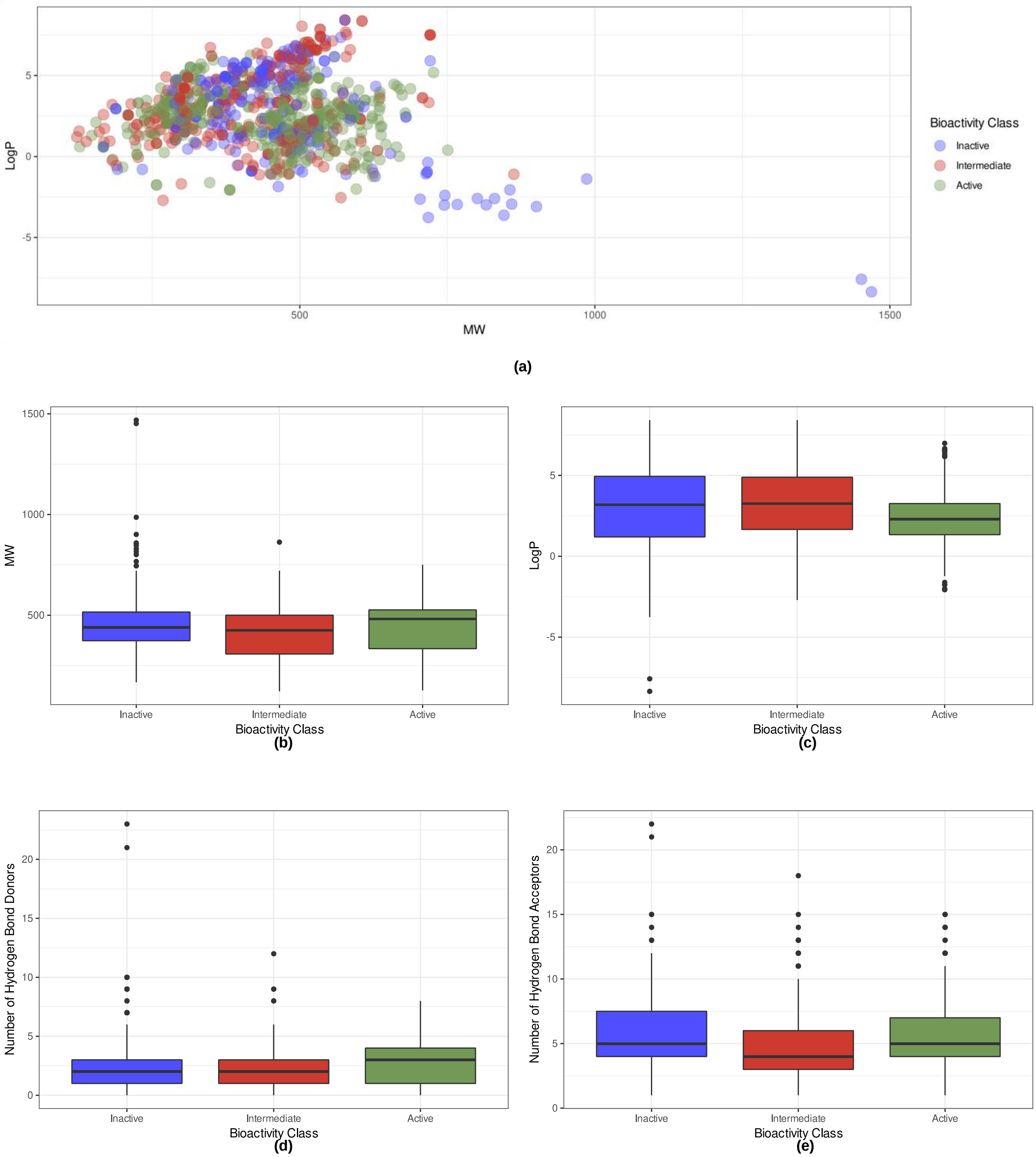
1. Full dataset, where the whole 881 features of the PubChem fingerprints as well as all bioactivity classes were used.
2. Without low variance features, the PubChem fingerprints that had insignificant variance were not included in the modeling and evaluation process.
3. Without the intermediate bioactivity class, only the compounds with IC_{50} that are less than or equal to 1000 nM and those greater than or equal to 10000 nM that included in the modeling and evaluation process.
4. Without low variance features and intermediate bioactivity class, which yielded the smallest dataset of all fours.

The modeling and evaluation part produced the performance dataset of the regression algorithms, with four dataset variances, where each scenario was repeated 100 times. The recorded performance parameters are R^2, Adjusted R^2, Root Mean Squared Error (RMSE), and Time Taken.

For the Modeling and Evaluation of the algorithms, we used a virtual machine with Ubuntu 20.04.6 with kernel 5.4.0-155-generic. The installed Python version is 3.8 along with libraries in used related to this research are lazypredict 0.2.12, numpy 1.24.3, and scikit-learn 1.3.0. The host has an Intel Xeon E5-2630 v4 CPU with eight cores allocated to the virtual machine. The allocated RAM is 32 GB.

C. STATISTICAL ANALYSIS

The statistical analysis part consists of two subparts. The first one is the chemical space analysis, where statistical methods were used to explore the characteristics of the compounds in each bioactivity class. This helps to understand the chemical nature of the compounds in each class. The other second subpart is statistical comparisons of the performance of each algorithm in each scenario (dataset variance). Descriptive and inference techniques are used to compare the algorithms. As the result could be erroneous due to a programming glitch in the library or the data characteristics that are not compatible with certain algorithms, the performance data must be filtered according to the theoretically possible R^2 and Adjusted R^2 values before applying the statistical techniques. Last, we draw a conclusion based on the statistical analysis results.



(a)

(b)

(c)

(d)

(e)

FIGURE 3. (a) Chemical space analysis as a function of molecular weight (MW) and octanol-water partition (LogP); (b)-(e) Distributions of the Lipinski's descriptors for each bioactivity class.

III. RESULT

A. CHEMICAL SPACE ANALYSIS

The chemical space analysis is aimed at understanding the chemical characteristics between the three bioactivity classes. First, the distributions of the bioactivity classes are visualized as the function of MW and LogP as shown in FIGURE 3(a). Second, the bioactivities are compared as the distributions of Lipinski's rule-of-five descriptors, as shown in FIGURE 3(b)-(e). The function of MW and LogP shows that most of the compounds are within a similar range of values, regardless of the bioactivity classes. There are few inactive compounds, and only a single red dot, denoting a single compound with

TABLE 1
Distribution normality test of the Lipinski's descriptors for each bioactivity class.

Class	Descriptor	Statistic	p	p.signif
inactive	LogP	0.951	<0.001	****
inactive	MW	0.814	<0.001	****
inactive	NumHAcceptors	0.867	<0.001	****
inactive	NumHDonors	0.689	<0.001	****
intermediate	LogP	0.988	0.007	**
intermediate	MW	0.985	0.001	**
intermediate	NumHAcceptors	0.869	<0.001	****
intermediate	NumHDonors	0.862	<0.001	****
active	LogP	0.992	0.003	**
active	MW	0.955	<0.001	****
active	NumHAcceptors	0.918	<0.001	****
active	NumHDonors	0.904	<0.001	****

TABLE 2
Results of the Kruskal-Wallis tests between bioactivity classes for each Lipinski's descriptor.

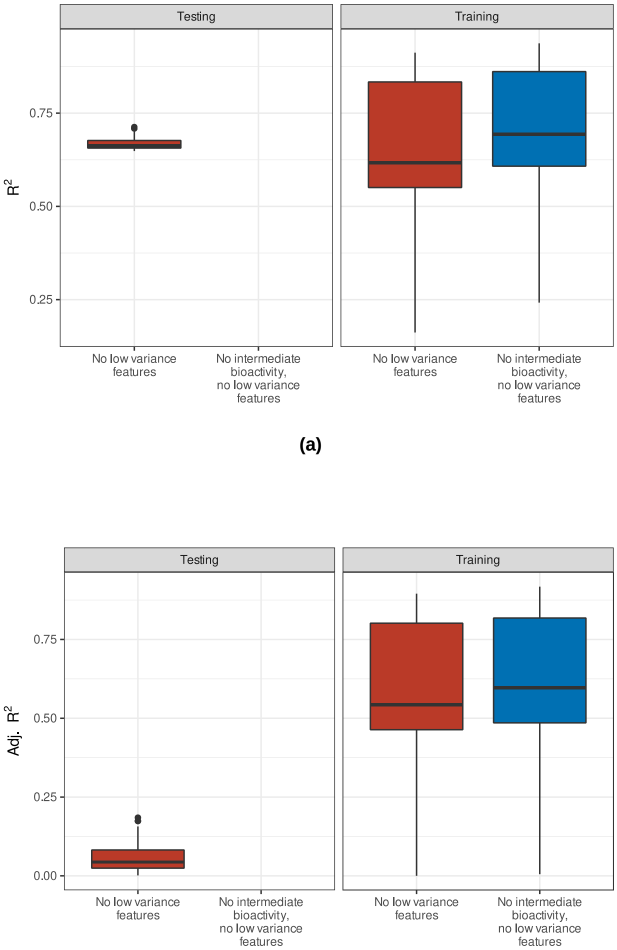
Descriptor	Statistics	df	p	p.signif
MW	1004.886	744	<0.001	****
LogP	1029.107	778	<0.001	****
NumHDonors	175.316	13	<0.001	****
NumHAcceptors	83.947	17	<0.001	****

intermediate bioactivity that strayed from the others. This fact indicates that most of the studied compounds are below 800 Da and LogP below 7.5. The boxplots of Lipinski's descriptors show that the compounds mostly satisfied Lipinski's rule of five.

T confirms the nonparametric natures of the descriptors in each bioactivity class. The Shapiro-Wilk's test of distribution was applied. Under a 95% confidence interval, all the p-values are below $\alpha = 0.05$, indicating the values are not normally distributed. Therefore, we continued with the Kruskal-Wallis test for each descriptor, using the bioactivity classes as groups, as shown in TABLE 2. As can be observed, there is at least a group of compounds that is significantly different. Hence, the analysis continued with a post-hoc test, they were using Dunn's method with Bonferroni adjustment. The results are shown in TABLE . It can be seen that in terms of molecular weight, the active and inactive compounds have no significant difference, while between the two and the intermediate class, there are significant differences. A similar pattern also can be observed in the number of hydrogen bond acceptors. In contrast, there is no significant difference in the number of hydrogen bond donors between the inactive and the intermediate compounds, while they are significant between the inactive and active compounds, and between the intermediate and active compounds.

B. PERFORMANCE OF THE REGRESSION ALGORITHMS

Before feeding the data into the algorithms, the dataset was assessed for its modelability. This step is required to ensure



(a)

(b)

FIGURE 4. **Distributions of valid R^2 and Adjusted R^2 values in each dataset variance.**

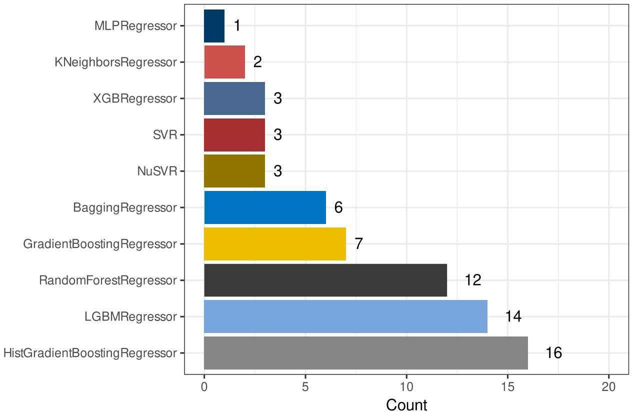


FIGURE 5. **Frequency of the experiment with valid R^2 and Adjusted R^2 values, grouped by algorithm. This only includes those trained and tested using the no low variance features dataset.**

TABLE 3
Results of the Dunn Posthoc Tests with Bonferroni adjustment between bioactivity classes for each Lipinski's descriptor.

Descriptor	group1	group2	n1	n2	statistic	p	p.adj	p.adj.signif
MW	inactive	intermediate	219	342	-3.149	0.001	0.003	**
	inactive	active	219	577	0.475	0.634	0.634	ns
	intermediate	active	342	577	4.547	<0.001	<0.001	****
LogP	inactive	intermediate	219	342	1.493	0.135	0.135	ns
	inactive	active	219	577	-3.669	<0.001	<0.001	***
	intermediate	active	342	577	-6.161	<0.001	<0.001	****
NumHDonors	inactive	intermediate	219	342	-0.273	0.784	0.784	ns
	inactive	active	219	577	4.291	<0.001	<0.001	****
	intermediate	active	342	577	5.337	<0.001	<0.001	****
NumHAcceptors	inactive	intermediate	219	342	-4.620	<0.001	<0.001	****
	inactive	active	219	577	0.762	0.446	0.446	ns
	intermediate	active	342	577	6.745	<0.001	<0.001	****

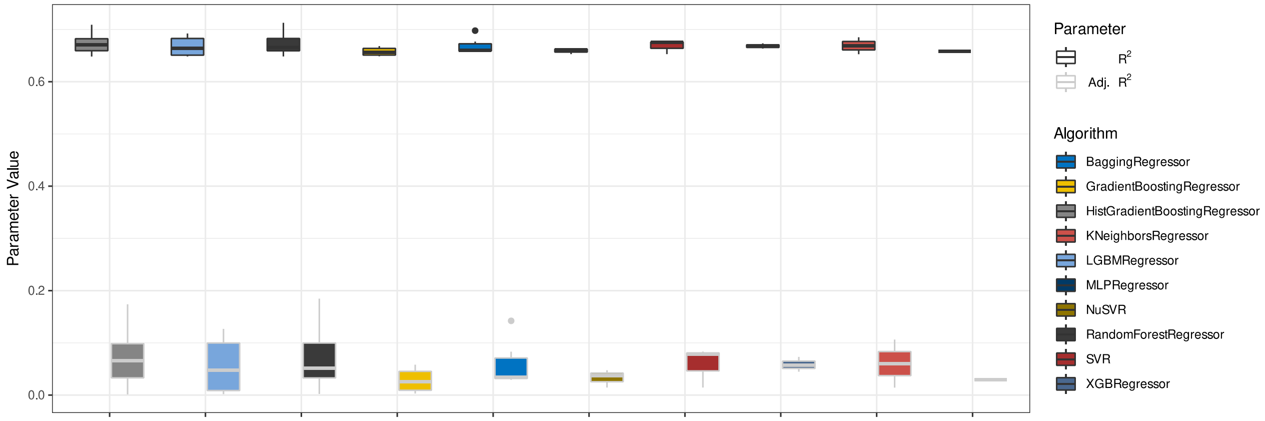


FIGURE 6. **Boxplot of the R^2 and Adjusted R^2 from model testing.**

the dataset could produce predictive QSAR models, identified by the Modelabiliy Index (MODI) that is greater than 0.65 [41]. In this study, the MODI is calculated utilizing an R code that was previously used in several published works by other researchers [2], [3]. The result shows that our curated dataset has a MODI of 0.768, indicating its modelability and the potential to yield predictive QSAR models.

1. R^2 AND ADJUSTED R^2

The QSAR modeling and evaluation part was executed with 42 algorithms, four dataset variances, and 100 repetitions, or in total, 33,600 experiments. However, due to various factors, the process only yielded 33,028 performance data. From the Exploratory Data Analysis (EDA) on the performance data, we found that much individual repetition resulted in invalid R^2 and/or Adjusted R^2 scores on the model testing part. Moreover, none of the performance parameters in the full dataset and the one without the intermediate bioactivity were found to be within the valid range. Fig. 4 shows the distributions of the valid R^2 and Adjusted R^2. It is only the dataset with omitted low variance features that have the testing performance parameters within the valid range. Therefore, further discussions are only concerned with the QSAR modeling yielded by this particular dataset. Out of 42 algorithms, only 10 fall within this category, with the data frequencies shown in Fig. 5.

Fig. 6 shows the values of the performance parameters from the model testing part. The shapes of the boxes correspond to the distributions of values. Since the Histogram Gradient Boosting Regressor (HGBR), Light Gradient Boosting Machine Regressor (LGBR), and Random Forest Regressor (RFR) are those that have more frequencies, therefore the boxes have larger spans, denoting the standard deviations. Therefore, further comparisons will be focused on these three algorithms. The Shapiro-Wilk tests of distribution normality in TABLE 4 show that the parameters produced by the LGBR algorithm are not normally distributed, hence further comparison must use a non-parametric method.

TABLE 4
Result of Shapiro-Wilk test on R2 and Adjusted R2 values of the algorithms with the top three frequencies

Algorithm	Parameter	Statistic	p	p.signif
HGBR	Adjusted R2	0.921	0.177	ns
HGBR	R^2	0.921	0.177	ns
LGBR	Adjusted R2	0.848	0.021	*
LGBR	R^2	0.848	0.021	*
RFR	Adjusted R2	0.866	0.059	ns
RFR	R^2	0.866	0.059	ns

TABLE 5
Results of the Kruskal-Wallis tests on R2 and Adjusted R2 between the algorithms with the top three frequencies.

Parameter	n	Statistic	df	p	p.signif
R^2	42	1.020	2	0.6	ns
Adjusted R2	42	1.020	2	0.6	Ns

TABLE 6
Results of the Shapiro-Wilk distribution normality test on the RMSE of the algorithms with the top three frequencies.

Algorithm	Statistic	p	p.signif
HGBR	0.947	0.447	ns
LGBR	0.984	0.993	ns
RFR	0.963	0.835	ns

TABLE 7
ANOVA test result on the RMSE of the algorithms with the top three frequencies.

Effect	DFn	DFd	F	p	ges	p.signif
Algorithm	2	39	1.048	0.36	0.051	ns

TABLE 8
Results of the Shapiro-Wilk distribution normality test on the Time Taken of the algorithms with the top three frequencies.

Phase	Algorithm	Statistic	p	p.signif
Training	HGBR	0.620	<0.001	****
	LGBR	0.878	0.054	ns
	RFR	0.958	0.759	ns
Testing	HGBR	0.952	0.001	**
	LGBR	0.647	<0.001	****
	RFR	0.922	<0.001	****

TABLE 9
Kruskal-Wallis test result on the time taken, grouped by the algorithms with the top frequencies, for training and testing phases.

Phase	n	statistic	df	p	p.signif
Testing	42	36.279	2	<0.001	****
Training	300	265.780	2	<0.001	****

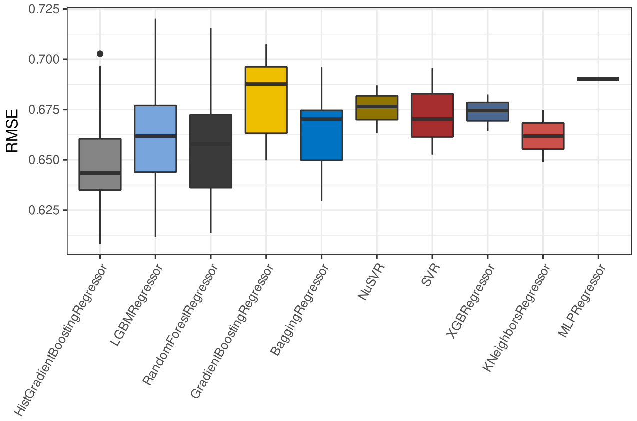


FIGURE 7. RMSE distribution of each algorithm in the testing stage.

TABLE 5 shows the results of the Kruskal-Wallis tests for both evaluation parameters between the algorithms. On either parameter, the difference is not significant.

2. COST FUNCTION

The Root Mean Squared Error (RMSE) was used as the cost function upon QSAR modeling and evaluation. It is calculated by taking the root of the squared average of the difference between each prediction and the actual value. Since the value reflects errors, which means the lower the RMSE, the better the model’s prediction capability. Fig. 7 shows the RMSE boxplots of the algorithms from the model testing of the non-low variance dataset. It can be seen that the top three frequencies tend to have lower RMSE compared to the other algorithms. The Shapiro-Wilk test in TABLE 6 shows that each of these top three frequency algorithms has normally distributed RMSE. Therefore, we continue comparing the RMSE using ANOVA. The result shown in TABLE 7 indicates that there is no significant difference in the RMSE of the HGBR, LGBR, and RFR algorithms.

3. TIME TAKEN

This parameter highlights the time needed for either training or testing. The lower the Time Taken means it took a shorter time for training and/or testing. Fig. 8 shows the distribution of time taken for training and testing the model produced by each algorithm. The Shapiro-Wilk test in TABLE 8 shows that for each algorithm in both phases, the durations are not normally distributed. In light of the distribution normality, the Kruskal-Wallis test is used to compare the time taken between the three algorithms. As shown in TABLE 9, there is at least one algorithm that has a significantly different duration in both training and testing. Using the Dunn test for post-hoc analysis shown in TABLE 10, we found the time taken between these three algorithms is significantly different. As reflected by the boxplots in Fig. 8, it is the LGBR that has the fastest time for both training and testing for this particular case.

IV. DISCUSSION

Conventional drug discovery is a tedious and expensive process. Fortunately, with the recent advances in computing technologies, in silico processes could be exercised for screening potential drug candidates. Classification is one of the supervised machine learning tasks that is gaining popularity for identifying potential bioactivity. In some studies, the bioactivity is based on the IC_{50} values that are grouped into two or three groups. Despite it is being a common approach, however, this practice reduces the information that can be learned by the model, as well as losing some statistical properties [34], [42], [43].

In this study, we applied regression algorithms to predict the IC_{50}, instead of classification. We argue that this approach should give more information, and better comparability with results from other studies. The SARS-CoV-2 Replicase Polyprotein 1ab is taken as a target. Using curated bioactivity data from the ChEMBL database, we tested the approach with 42 regression algorithms and four dataset variances that repeated 100 times. Counting the training and testing phases, the experiment yields 33,028 data or only about 98.29% of the expected 33,600 data. By investigating the experiment logs,

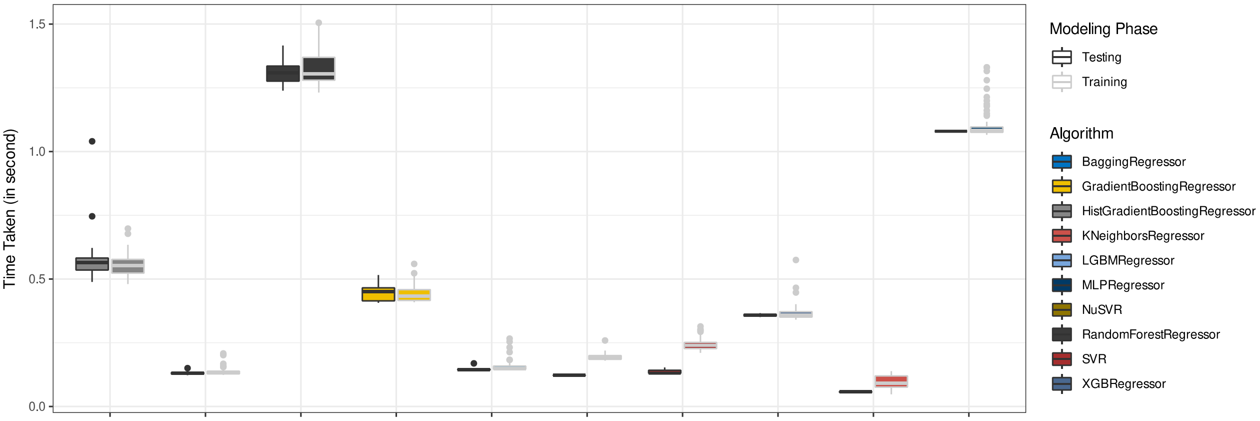


FIGURE 8. Distributions of the training and testing durations of each algorithm.

TABLE 10
Results of the Dunn test with Bonferroni adjustment for the time taken between algorithms.

Phase	group1	group2	n1	n2	Statistic	p	p.adj	p.adj.signif
Testing	HGBR	LGBR	16	14	-3.341	<0.001	0.001	**
	HGBR	RFR	16	12	2.988	0.002	0.002	**
	LGBR	RFR	14	12	6.008	<0.001	<0.001	****
Training	HGBR	LGBR	100	100	-8.151	<0.001	<0.001	****
	HGBR	RFR	100	100	8.151	<0.001	<0.001	****
	LGBR	RFR	100	100	16.302	<0.001	<0.001	****

the reduction was caused by some algorithms that failed to run properly due to glitches or incompatibilities with the data. Moreover, much of the resulting testing phase performance parameters, namely the R^2 and the Adjusted R^2 are falling out of the theoretically possible range. In the end, only some of the results from the no-low variance dataset can have meaningful model testing performance parameters. Algorithms-wise, only 10 out of 42 satisfy the requirement, and only three of them can be considered stable enough for inference, judging from the number of valid R^2 and Adjusted R^2 values.

From these three algorithms, namely Histogram Gradient Boosting Regression (HGBR), Light Gradient Boosting Machine Regression (LGBR), and Random Forest Regression (RFR), they share similar qualities in terms of predicting capability (R^2 and Adjusted R^2) and the observed cost function (RMSE). It is only the Time Taken that has significant differences between the algorithms, where LGBR came up with the shortest durations, followed by HGBR and then the RFR.

Random Forest Classifier is a known Machine Learning method in drug discovery, as demonstrated in several studies [2]–[6]. Despite the popularity of its counterpart, as far as this study is concerned, the performance of the Random Forest algorithm can be matched by HGBR and LGBR. If there is an emphasis on time, the LGBR should be highly considered, although further investigations must be carried out. On the contrary, the Support Vector Regression (SVR) did not perform as expected, despite it being reported to be proficient in QSAR modeling [7].

Our study has several implications for the use of Machine Learning algorithms in drug discovery. First, we have explored the use of regression algorithms where classification is the commonly applied technique in the field. We have also found that the application of regression algorithms in this particular field faces several computational as well as algorithmic challenges where most of the model performance parameters in the testing phase were invalid. However, the possibility of getting a higher range of information compared to the discretized approach in classification becomes a strong point for more investigation. Despite our experiment yielding HGBR, LGBR, and RFR as the most stable regression algorithms that might not be the case with other datasets, such as different targets or molecular descriptors. Moreover, compared to the accuracy of the previous study that targeted the SARS-CoV-2 3CLpro [3], which is about 70%, the algorithms we evaluated still performed poorly. Whether hyperparameter tuning can be exploited to increase performance, should be thoroughly explored. Therefore, further studies must be undertaken to answer these questions.

V. CONCLUSION

Drug discovery is a long and expensive process. Fortunately, due to the advancement in computational technology and cheminformatics, those properties could be reduced for good reasons. When to COVID-19 pandemic was starting to hit the world, there was an urgency for an accelerated development of novel as well as repurposed drugs. Machine Learning which is a branch of computational science is extensively used in this part, with the classification task as the major approach.

In this study, we explored the use of regression algorithms to predict the compounds' inhibitory bioactivity toward the SARS-CoV-2 replicase polyprotein 1ab. By using the IC_{50} data from the ChEMBL database and PubChem descriptor for converting the compounds into numeric vectors, we experimented with the combinations of algorithms and dataset variances. Evaluating the results, we uncovered that the use of regression algorithms is challenging, for we got many invalid performance parameters from the model testing phase, where the R^2 and/or Adjusted R^2 values are out of the theoretical range. However, the use of regression algorithms in this particular field would yield more information than the discretized ones in the classification task. We found that for the particular dataset used in this study, the regression algorithms of Histogram Gradient Boosting, Light Gradient Boosting Machine, and Random Forest yielded the highest counts of valid results.

Regardless of the use of Random Forest Classification in some previous studies, from our experiments, its regression counterpart has no statistically significant performance differences compared to the Histogram Gradient Boosting Regressor and the Light Gradient Boosting Machine Regressor. Surprisingly, the Light Gradient Boosting Machine Regressor has a significantly lower duration for either training or testing, making it a promising algorithm to be explored for this particular case.

There should be more studies that investigate the use of regression algorithms in predicting inhibitory bioactivity. The three algorithms should even be explored from several different approaches. The performances of each algorithm could be improved by using the appropriate hyperparameter tuning. In this study, we only converted the dataset to a feature matrix with a single molecular fingerprint. As different fingerprint leads to a different form of feature matrix, it should lead to different performance. This could be addressed in future studies. Moreover, a different target enzyme or protein might have another set of algorithms that suits it. Hence, this study then could be repeated with a different target.

REFERENCES

[1] M. Davies *et al.*, “ChEMBL web services: streamlining access to drug discovery data and utilities,” *Nucleic Acids Res.*, vol. 43, no. W1, pp. W612–W620, Jul. 2015, doi: 10.1093/nar/gkv352.

[2] A. A. Malik, C. Phanus-umporn, N. Schaduangrat, W. Shoombuatong, C. Isarankura-Na-Ayudhya, and C. Nantasenamat, “HCVpred: A web server for predicting the bioactivity of hepatitis C virus NS5B inhibitors,” *J. Comput. Chem.*, vol. 41, no. 20, pp. 1820–1834, Jul. 2020, doi: 10.1002/JCC.26223.

[3] N. Ferdous *et al.*, “Mpropred: A machine learning (ML) driven Web-App for bioactivity prediction of SARS-CoV-2 main protease (Mpro) antagonists,” *PLoS One*, vol. 18, no. 6 June, pp. 1–21, 2023, doi: 10.1371/journal.pone.0287179.

[4] T. Lerksuthirat, S. Chitphuk, W. Stitchantrakul, D. Dejsuphong, A. A. Malik, and C. Nantasenamat, “Parp1Pred: a Web Server for Screening the Bioactivity of Inhibitors Against Dna Repair Enzyme Parp-1,” *EXCLI J.*, vol. 22, pp. 84–107, 2023, doi: 10.17179/excli2022-5602.

[5] N. S. Ramadhanti, W. A. Kusuma, I. Batubara, and R. Heryanto, “Random Forest to Predict Eucalyptus as a Potential Herb in Preventing Covid19,” in *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Oct. 2021, pp. 01–05, doi: 10.1109/CIBCB49929.2021.9562940.

[6] V. Svetnik, A. Liaw, C. Tong, J. Christopher Culberson, R. P. Sheridan, and B. P. Feuston, “Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling,” *J. Chem. Inf. Comput. Sci.*, vol. 43, no. 6, pp. 1947–1958, Nov. 2003, doi: 10.1021/CI034160G/SUPPL_FILE/CI034160GSI20031008_041202.ZIP.

[7] R. Rodríguez-Pérez and J. Bajorath, “Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery,” *J. Comput. Aided. Mol. Des.*, vol. 36, no. 5, pp. 355–362, May 2022, doi: 10.1007/S10822-022-00442-9/FIGURES/5.

[8] A. Mullard, “Biotech R&D spend jumps by more than 15,” *Nat. Rev. Drug Discov.*, vol. 15, no. 7, p. 447, 2016, doi: 10.1038/nrd.2016.135.

[9] D. E. Salazar and G. Gormley, *Modern Drug Discovery and Development*, vol. 26. Elsevier Inc., 2017.

[10] S. Mishra, “Artificial Intelligence: A Review of Progress and Prospects in Medicine and Healthcare,” *J. Electron. Electromed. Eng. Med. Informatics*, vol. 4, no. 1, pp. 1–23, 2022, doi: 10.35882/jeeemi.v4i1.1.

[11] D. Cucinotta and M. Vanelli, “WHO declares COVID-19 a pandemic,” *Acta Biomedica*, vol. 91, no. 1. Mattioli 1885, pp. 157–160, 2020, doi: 10.23750/abm.v91i1.9397.

[12] Trie Maya Kadarina and R. Priambodo, “Performance Evaluation of IoT-based SpO2 Monitoring Systems for COVID-19 Patients,” *J. Electron. Electromed. Eng. Med. Informatics*, vol. 3, no. 2, pp. 64–71, Jul. 2021, doi: 10.35882/JEEEMI.V3I2.1.

[13] G. Li and E. De Clercq, “Therapeutic options for the 2019 novel coronavirus (2019-nCoV),” *Nat. Rev. Drug Discov.*, vol. 19, no. 3, pp. 149–150, 2020, doi: 10.1038/d41573-020-00016-0.

[14] V. T. Sabe *et al.*, “Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review,” *Eur. J. Med. Chem.*, vol. 224, p. 113705, Nov. 2021, doi: 10.1016/j.ejmech.2021.113705.

[15] F. Yang *et al.*, “Machine Learning Applications in Drug Repurposing,” *Interdiscip. Sci. – Comput. Life Sci.*, vol. 14, no. 1, pp. 15–21, 2022, doi: 10.1007/s12539-021-00487-8.

[16] F. E. Agamah *et al.*, “Computational/in silico methods in drug target and lead prediction,” *Brief. Bioinform.*, vol. 21, no. 5, pp. 1663–1675, Sep. 2020, doi: 10.1093/BIB/BBZ103.

[17] P. Subhaswaraj and B. Siddhardha, “Molecular docking and molecular dynamic simulation approaches for drug development and repurposing of drugs for severe acute respiratory syndrome-Coronavirus-2,” *Comput. Approaches Nov. Ther. Diagnostic Des. to Mitigate SARS-CoV2 Infect. Revolut. Strateg. to Combat Pandemics*, pp. 207–246, Jan. 2022, doi: 10.1016/B978-0-323-91172-6.00007-8.

[18] A. S. Omrani *et al.*, “Ribavirin and interferon alfa-2a for severe Middle East respiratory syndrome coronavirus infection: a retrospective cohort study,” *Lancet. Infect. Dis.*, vol. 14, no. 11, pp. 1090–1095, Nov. 2014, doi: 10.1016/S1473-3099(14)70920-X.

[19] W. Yan, Y. Zheng, X. Zeng, B. He, and W. Cheng, “Structural biology of SARS-CoV-2: open the door for novel therapies,” *Signal Transduct. Target. Ther. 2022 71*, vol. 7, no. 1, pp. 1–28, Jan. 2022, doi: 10.1038/s41392-022-00884-5.

[20] V. Mody *et al.*, “Identification of 3-chymotrypsin like protease (3CLPro) inhibitors as potential anti-SARS-CoV-2 agents,” *Commun. Biol. 2021 41*, vol. 4, no. 1, pp. 1–10, Jan. 2021, doi: 10.1038/s42003-020-01577-x.

[21] D. Shaji, S. Yamamoto, R. Saito, R. Suzuki, S. Nakamura, and N. Kurita, “Proposal of novel natural inhibitors of severe acute respiratory syndrome coronavirus 2 main protease: Molecular docking and ab initio fragment molecular orbital calculations,” *Biophys. Chem.*, vol. 275, Aug. 2021, doi: 10.1016/j.bpc.2021.106608.

[22] R. Alexpandi, J. F. De Mesquita, S. K. Pandian, and A. V. Ravi, “Quinolines-Based SARS-CoV-2 3CLpro and RdRp Inhibitors and Spike-RBD-ACE2 Inhibitor for Drug-Repurposing Against COVID-19: An in silico Analysis,” *Front. Microbiol.*, vol. 11, Jul. 2020, doi:

10.3389/fmicb.2020.01796.

[23] L. Erlina *et al.*, “Virtual screening of Indonesian herbal compounds as COVID-19 supportive therapy: machine learning and pharmacophore modeling approaches,” *BMC Complement. Med. Ther.*, vol. 22, no. 1, p. 207, Dec. 2022, doi: 10.1186/s12906-022-03686-y.

[24] A. Kumar, S. Loharch, S. Kumar, R. P. Ringe, and R. Parkesh, “Exploiting cheminformatic and machine learning to navigate the available chemical space of potential small molecule inhibitors of SARS-CoV-2,” *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 424–438, 2021, doi: 10.1016/j.csbj.2020.12.028.

[25] K. Mohamed, N. Yazdanpanah, A. Saghazadeh, and N. Rezaei, “Computational drug discovery and repurposing for the treatment of COVID-19: A systematic review,” *Bioorg. Chem.*, vol. 106, p. 104490, Jan. 2021, doi: 10.1016/J.BIOORG.2020.104490.

[26] X. Huang, R. Pearce, G. S. Omenn, and Y. Zhang, “Identification of 13 Guanidinobenzoyl- or Aminidinobenzoyl-Containing Drugs to Potentially Inhibit TMPRSS2 for COVID-19 Treatment,” *Int. J. Mol. Sci.*, vol. 22, no. 13, p. 7060, Jun. 2021, doi: 10.3390/ijms22137060.

[27] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science (80-.).*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[28] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, and M. J. Ahsan, “Machine Learning in Drug Discovery: A Review,” *Artif. Intell. Rev. 2021 553*, vol. 55, no. 3, pp. 1947–1999, Aug. 2021, doi: 10.1007/S10462-021-10058-4.

[29] F. Sulistiawan, W. A. Kusuma, N. S. Ramadhanti, and A. Tedjo, “Drug-Target Interaction Prediction in Coronavirus Disease 2019 Case Using Deep Semi-Supervised Learning Model,” in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Oct. 2020, pp. 83–88, doi: 10.1109/ICACSIS51025.2020.9263241.

[30] S. Aini, W. A. Kusuma, M. K. D. Hardhienata, and Mushthofa, “Network-Based Molecular Features Selection to Predict the Drug Synergy in Cancer Cells,” *J. Electron. Electromed. Eng. Med. Informatics*, vol. 5, no. 3, pp. 168–176, 2023, doi: 10.35882/jeemi.v5i3.307.

[31] T. Yu *et al.*, “Exploring the Chemical Space of CYP17A1 Inhibitors Using Cheminformatics and Machine Learning,” *Molecules*, vol. 28, no. 4, pp. 1–23, 2023, doi: 10.3390/molecules28041679.

[32] A. Fadli *et al.*, “Screening of Potential Indonesia Herbal Compounds Based on Multi-Label Classification for 2019 Coronavirus Disease,” *Big Data Cogn. Comput. 2021, Vol. 5, Page 75*, vol. 5, no. 4, p. 75, Dec. 2021, doi: 10.3390/BDCC5040075.

[33] G. W. Caldwell, Z. Yan, W. Lang, and J. A. Masucci, “The IC50 Concept Revisited,” *Curr. Top. Med. Chem.*, vol. 12, no. 11, pp. 1282–1290, May 2012, doi: 10.2174/156802612800672844.

[34] C. Bennette and A. Vickers, “Against quantiles: Categorization of continuous variables in epidemiologic research, and its discontents,” *BMC Med. Res. Methodol.*, vol. 12, no. 1, pp. 1–5, Feb. 2012, doi: 10.1186/1471-2288-12-21/FIGURES/3.

[35] D. F. Sengkey, A. Jacobus, and F. J. Manoppo, “Effects of kernels and the proportion of training data on the accuracy of SVM sentiment analysis in lecturer evaluation,” *IAES Int. J. Artif. Intell.*, vol. 9, no. 4, Dec. 2020, doi: 10.11591/IJAI.V9.I4.PP%P.

[36] “RDKit: Open-source cheminformatics.” https://www.rdkit.org/ (accessed Oct. 25, 2023).

[37] “rdkit/rdkit: 2023_03_1 (Q1 2023) Release,” doi: 10.5281/ZENODO.7880616.

[38] C. W. Yap, “PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints,” *J. Comput. Chem.*, vol. 32, no. 7, pp. 1466–1474, May 2011, doi: 10.1002/JCC.21707.

[39] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011, Accessed: Oct. 26, 2023. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html.

[40] S. R. Pandala, “shankarpandala/lazypredict: Lazy Predict help build a lot of basic models without much code and helps understand which models works better without any parameter tuning.” https://github.com/shankarpandala/lazypredict (accessed Oct. 26, 2023).

[41] A. Golbraikh, E. Muratov, D. Fourches, and A. Tropsha, “Data Set Modelability by QSAR,” *J. Chem. Inf. Model.*, vol. 54, no. 1, pp. 1–4, Jan. 2014, doi: 10.1021/ci400572x.

[42] K. Felsenstein and K. Pötzelberger, “The Asymptotic Loss of Information for Grouped Data,” *J. Multivar. Anal.*, vol. 67, no. 1, pp. 99–127, Oct. 1998, doi: 10.1006/jmva.1998.1759.

[43] K. E. Markon, M. Chmielewski, and C. J. Miller, “The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review.,” *Psychol. Bull.*, vol. 137, no. 5, pp. 856–879, Sep. 2011, doi: 10.1037/a0023678.

BIOGRAPHY

DANIEL FEBRIAN SENGKEY is an Assistant Professor at the Undergraduate Program in Informatics, Department of Electrical Engineering, Faculty of Engineering, Universitas Sam Ratulangi, Manado-Indonesia. He graduated from the Undergraduate Program in Electrical Engineering of the same department in 2012. Later in 2015, he achieved his Master of Engineering degree from the Master Program in Electrical Engineering, under the Information Technology concentration, in the Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta-Indonesia. His current research interest is in the implementation of Machine Learning in various fields, with a focus on Bioinformatics.

ANGELINA STEVANY REGINA MASENGI is currently the Acting Secretary of the Department of Clinical Pharmacology and Therapy, Faculty of Medicine, Universitas Sam Ratulangi. She achieved her Bachelor of Medicine and Medical Doctor profession from the Faculty of Medicine, Universitas Pelita Harapan in 2008 and 2010, respectively. She holds a master's degree in Biomedics, achieved in 2016 from the Master's Program in Biomedical Science, at Universitas Indonesia. Since 2018, she has been a tenured lecturer at Universitas Sam Ratulangi. Despite her assignment in the department, she is participating actively in teaching activities at several undergraduate programs, namely: Medicine, Nursing, Dentistry, and Pharmacy. She was also a member of the teaching team of the Bioinformatics course, in the Undergraduate Program in Informatics.