Random Forest to Predict Eucalyptus as a Potential Herb in Preventing Covid19

Nabila Sekar Ramadhanti
Tropical Biopharmaca Research Center, IPB University
Bogor, Indonesia
nabila.s.ramadhanti@gmail.com

Wisnu Ananta Kusuma
Department of Computer Science and
Tropical Biopharmaca Research Center, IPB University
Bogor, Indonesia
ananta@apps.ipb.ac.id

Irmanida Batubara
Department of Chemistry and
Tropical Biopharmaca Research Center, IPB University
Bogor, Indonesia
ime@apps.ipb.ac.id

Rudi Heryanto
Department of Chemistry and
Tropical Biopharmaca Research Center, IPB University
Bogor, Indonesia
rudi_heryanto@apps.ipb.ac.id

2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) | 978-1-6654-0112-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/CIBCB49929.2021.9562940

***Abstract*—The covid-19 pandemic had been on the rise since the beginning of 2020. In Indonesia itself, the first case was identified on 3rd March 2020, then peaked at around the end of January 2021. Even though the recent number of covid-19 cases is not as much as the peak time, the positive case has been increasing from around 2600 to 6300 cases every day in the last month. This phenomenon is urging people to take better care of their health. One of the alternatives Indonesian takes to maintain and increase their health is using herbal medicine. Indonesia is one of the countries with a flourishing number of herbal species. Eucalyptus is one of herbal plants with lots of benefits. Even before the pandemic eucalyptus oil has been used for daily use by many in Indonesia. In this study, we predict the compounds in eucalyptus which have any interaction with protein in SARS-COV-2 virus using machine learning method, namely Random Forest. This is one of the applications of the drug-discovery method, drug repurposing, which used existing drug-target interaction data as a model to predict drug compounds with unidentified interaction with targets. Applying this method, we predicted some compounds found in eucalyptus, such as alpha-terpinene, and 1,8-cineole might have an interaction with covid-19 protein thus eucalyptus can be used as a preventive measure.**

Keywords—Covid19, Eucalyptus, Herbal, Machine learning, Random Forest

I. Introduction

The covid-19 pandemic had been on the rise since the beginning of 2020. In Indonesia itself, the first case was identified on 3rd March 2020, then peaked at around the end of January 2021. Even though the recent number of covid-19 cases is not as much as the peak time, the positive case has been increasing from around 2600 to 6300 cases every day in the last month. There have been lots of efforts to contain the pandemic, such as health protocol in public places, lockdown, and vaccination. This phenomenon is raising people awareness to take better care of their health. One of the alternatives Indonesian takes to maintain and increase their health is using herbal medicine. Indonesia is one of the countries with a flourishing number of herbal species.

Eucalyptus is one of herbal plants with lots of benefits. Even before the pandemic eucalyptus oil has been used for daily use by many in Indonesia. In order to add more scientific background in herbal medicine, in accordance to the regulation of the Minister of Health of the Republic of Indonesia No.003/2010 in herbal medicine (Jamu) scientification, in this study, we predict the compounds in Eucalyptus which have any interaction in covid19 protein using machine learning, to be exact, Random Forest. This is one of the applications of the drug-discovery method, drug repurposing, which used existing drug-target interaction data as a model to predict drug compounds with unidentified interaction with targets.

Drug repurposing has already been talked about in lots of older publication, dated in year 2011 [1], talking about drug repurposing from academic perspective and so on, even in 2019, this method still seemed to be in infant stage [2] with a great potential. In most recent year, there are lots of many published papers about it, but [3] stated the potential of drug repurposing for Covid19.

Machine learning has proved to be able to improve drug repurposing method. Support vector machine (SVM), one of a basic machine learning method able to improve docking score for direct inhibitors of *Mycobacterium tuberculosis* [4]. Another study [5] build a webserver based on G-protein-coupled receptors (GPCRs) drug interactions predictor using RF classifier on its initial prediction. Combining SVM, random forest (RF) and multilayer perceptron (MLP), and a pharmacophore modelling [6] created a model to predict Covid19 potential drugs and predict *Psidium guajava* (guava) as a potential antiviral candidate. Applying RF as a base model with some sampling technique to balance the training data, in this study, we predict compounds found in eucalyptus to Covid19 protein targets.

II. Random Forest

Random forest is a tree-based classifier, which essentially a machine learning method, in which, a classifier consisted of a number of decision trees $\{h_k(x, T_k)\}$, $k = 1, 2, \ldots, L$, where T_k are formed from a random vector sampled from the input, and the most popular class from the trees are chosen to be the class for random data x [7]. The illustration of random forest can be seen on Fig.5.

Random forest produces multiple decision trees using a random number of samples and features by uniting them in a classifier [8]. Random forest has some advantages to use in drug-target interaction data, as [9] stated it's good at prediction when most variable are noise and does not overfit.

978-1-6654-0112-8/21/$31.00 ©2021 IEEE

The data used here indicate that most of the unidentified interaction is indeed unidentified, so it fits as a noise, and might be a good fit to use Random Forest. Some research also used Random Forest Classifier in early stage of drug repositioning, [10] used Random Forest for drug-target interaction prediction with a gold standard dataset and performed 0.91 area under the ROC curve score.

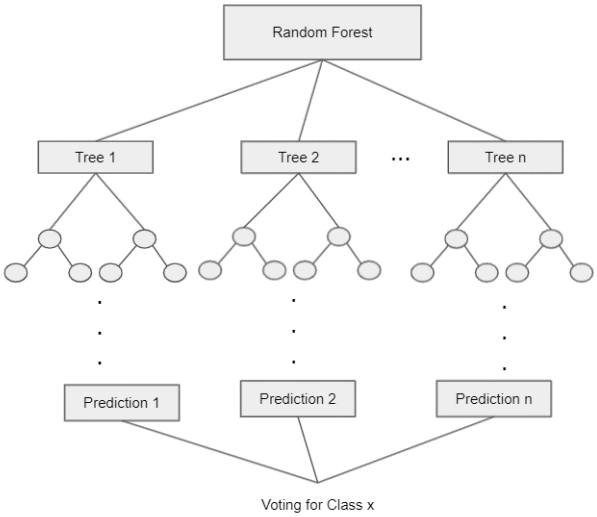

Fig. 1. Random forest illustration

Random forest classifier uses Gini Index to filter the attribute. For example, a training set T, and a random data in class Ci, Gini index can be written as in equation (1):

$$\sum\sum_{j\neq i}(f(C_i,T)\,/|T||(C_j,T()|T|)| \quad (1)$$

where f(Ci,T)/|T| is the probability that the data belongs to class Ci. [11]

As stated in [12], bagging classifier might be used to enhance performance. The Random Forest used in this research referring to a baggingClassifier in sklearn which, pretty much the same, the only difference, in sklearn baggingClassifier, there is a parameter we need to use, in which to set the weight of the class as a balanced class to evade the impact of the imbalanced data we used.

III. METHOD

A. Data Acquisition

The initial data used in this research was acquired in other studies [13][14] in 2020. [13] predicted the structure of SARS-CoV-2 gene using homology modelling. [14] stated some potential drugs based on a significant binding affinity score on drug-target interaction. There were 81 virus-based drugs and 17 human-based drugs, in total resulting in 98 and 23 initial drugs and proteins.

From these drugs and proteins, we explore more data by scraping SuperTarget web [15]. The proteins are used to find more compounds interacting with proteins, while the drugs are used to find more targets interacting with drugs. From this process we acquire a total of 675 interactions, in which consist of 119 drug compounds and 335 targets. The total of possible interaction is 119 drugs*335 targets = 39 865 interactions. This data was the same data used in earlier study [6] which infer Psidium guajava (guava) as one of Indonesia's commodities with potential to be an antiviral. Furthermore we also used eucalyptus compound from Table I and covid19 protein targets from [6] to predict eucalyptus as a potential covid19 antiviral as described before.

TABLE I. EUCALYPTUS COMPOUND

Compound	Pubchem CID	Compound	Pubchem CID
1,8-Cineol	2758	β-guaiene	15560252
α-terpinolene (C10H16)	7462	Cubenol (C15H26O)	519857
Caryophyllene	5281515	α-cadinol (C15H26O)	10398656
γ-terpinene (C10H16)	7461	α-eudesmol (C15H26O)	92762
α-amorphene (C15H24)	12306052	Bulnesol (C15H26O)	90785
Cadina 3,9-diena	10657	1,4-dimethoxy 6,7,8,9	613233
Germacrene B (C15H24)	5281519	1,4-naphtalenedione (C10H6O2)	8530
Guaiol (C15H26O)	227829		

The feature used as the data descriptor was also the same as [6] using PubChem fingerprint as drug compound descriptor and dipeptide as protein target descriptor, more detailed illustration about the data used is shown on Fig.1. Every compound with interaction found from either paper or SuperTarget is indicated as having a line bridging between the compound and protein.

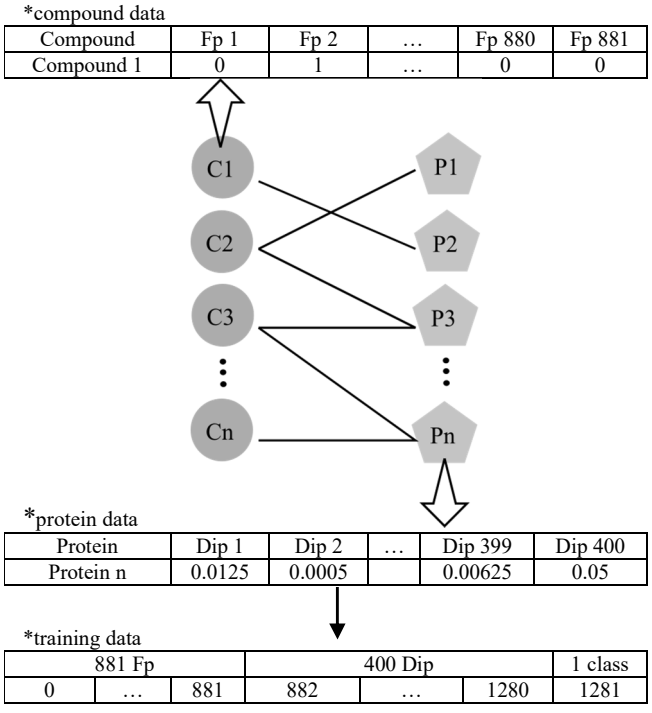

Fig. 2. Illustration of data used in this study. (Fp=Fingerprints, Dip=Dipeptide composition descriptors)

Compound fingerprint was obtained from encoding of Simplified Molecular-Input Line-Entry System (SMILES), which is one kind of representation of compound [16]. On this study, to retrieve the fingerprint, we used PubChemPy, a python package to help interact with PubChem database. PubChem fingerprint has 881 features represented by binary number (0/1) that helps accelerate machine learning process because of its size which only need 1 bit of space per feature.

The sub-structure used on PubChem fingerprint is based on 2D-structure of a compound that is used on similarity search [6] which is in accordance with our purpose to find similar compound in eucalyptus from synthetic drug-target interaction data.

Protein in this study is represented in dipeptide composition, which is a set of combinations of two amino acid composition. Amino acid composition in a protein is consisted of 20 components, each symbolized by a single letter. Amino acid composition itself can be a descriptor, but we used dipeptide composition to cover amino acid composition disadvantage as it cannot convey the order of a sequence [17]. Dipeptide composition have already been used in [6] and obtained a good result for a classification or prediction problem.

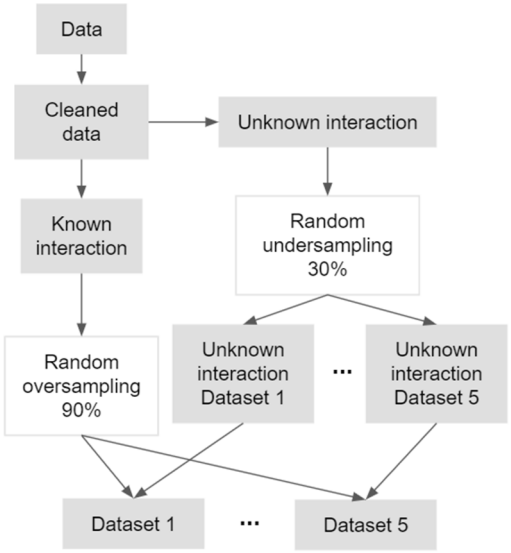

Fig. 3. Data balancing process

B. Balancing Data

The complete data from the data acquisition process was imbalance with 1:57 ratio, to be exact, 685 positive class, which have interaction between the compounds and proteins, and 39 180 combinations of unlabelled compound and protein data. This condition might affect the machine learning process to lean more to the dominating class, in this case, the unidentified interactions. A data balancing process needs to be done to avoid full effect of the imbalanced data.

First step into balancing data, we used the Scikit-learn [18] model_selection to randomly split the unidentified interaction data (data with label 0) into 3:7 data. After we reduced the unidentified data, we also oversampled the data using oversample in imblearn Scikit-learn. The 30% of the data (11 754) was used as a base to oversample data with identified interaction (data with label 1) for 90% of the amount. After this step, we got 10:9 in total of 22 332 data for unidentified and identified interactions which visualized in Fig.2. To get a more precise result, we ran through this step to get five datasets randomly split and oversampled, more structured steps we did here can be seen on Fig.4.

C. Building Prediction Model

For this research, we performed Random Forest in five different dataset and combine the result from each dataset. The five balanced datasets from the last process were tuned to get the best variable whilst applying the machine learning method with GridSearchCV function from Scikit-learn. The visualization of this process can be seen on Fig.5.

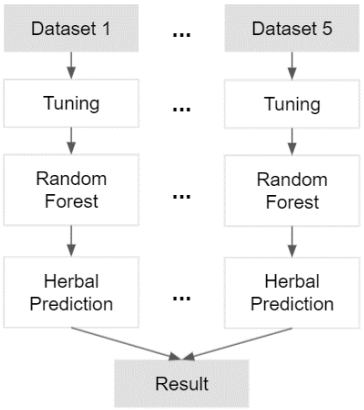

Fig. 5. Prediction process

Table II shows the variables tested on the tuning process, max_features and max_samples, max_features adjust the number of features, while max_samples adjust the number of samples to be used in each tree. Instead of directly using Random Forest Classifier, we used Bagging Classifier with

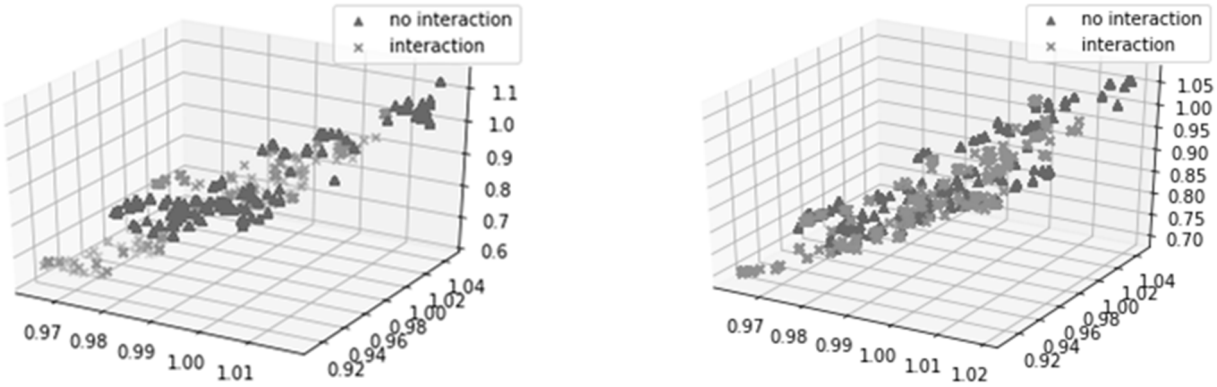

Fig. 4. Data distribution before and after balancing process

decision tree classifier as a base estimator and balanced weight class. From the tuning process, we got five different Random Forest models with 500 as the best max_features and 15 000 as the best max_samples.

TABLE II. TUNED VARIABLE

Parameter	Variable
max_features	100, 300, 500, 800, 1000, 1281
max_samples	1000, 5000, 10 000 15 000, 20 000

The GridSearchCV then was set to calculate the score with five-fold cross-validation to evaluate the score of the models. The area under the ROC curve (AUC), f-measure, precision, recall, and accuracy score averaged at 0.999661, 0.988228, 0.976775, 1, and 0.98871. If you just look at the recall and accuracy score, it might seem too overfit, but this score is also supported by the AUC and f-measure score.

IV. RESULT AND DISCUSSION

The scores in Table III means that the models generated are good enough to be a prediction model. We can see that the accuracy and F-measure score is not so different, from these scores, we can say that there was no bias on the performance of the models. The high score of AUC and F-measure also proves that the balancing method we use was able to overcome the imbalanced data problems we had.

TABLE III. SCORE

Model	AUC	F-measure	Precision	Recall	Accuracy
1	0.999679	0.988327	0.977108	1	0.98881
2	0.999711	0.989872	0.979951	1	0.990305
3	0.999555	0.98679	0.97393	1	0.987315
5	0.999754	0.988472	0.977216	1	0.988946
4	0.999604	0.987679	0.975671	1	0.988176

We predicted eucalyptus compounds, shown in Table I with covid19 targets [6] based on the five Random Forest models. The compound - target couple would be considered as having any interaction if the probabilities score is >=0.5. Fig. 6 shown that whilst none of the compounds have any interaction with Spike, most of the compounds have a high probability to have any interaction with other targets such as main protease (3CLPro), Papain-like protease (PLPro) and RNA-dependent RNA polymerase (RdRp), except 1,8-cineole on RdRp.

1,8-cineole has been used in some drugs as anti-inflammatory compounds, but the latest studies imply the compound is also a potential antiviral. [19] reported about the effect of essential oil on Bovine Viral Diarrhea Virus, which include 1,8-cineole as one of the components. Meanwhile [20] studying essential oil for Covid19 therapy also imply the efficacy of 1,8-cineole in Sars-Cov-2 proteinase, but mostly used to ease many kinds of respiratory ailments. [21] also imply that some compounds of aromatic plants and essential oils, including 1,8-cineole, could act as inhibitor of Covid19.

However, this study is a computational work that focuses more on screening in order to minimize the search space for finding potential herbal compound candidates. In the drug discovery stage, this study must be followed up with in vitro, in vivo and clinical trials. This series of tests is needed to determine the efficacy and safety of these herbal compounds. In addition, further research is needed to determine the right dose for prevention and treatment. It will be even better if a metabolomic analysis is carried out to determine with more precision the metabolites that play role in this inhibitory mechanism.

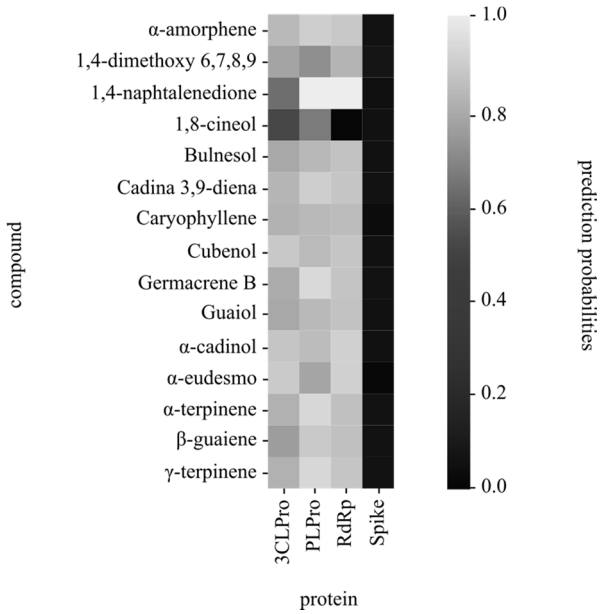

Fig. 6. Prediction probabilities result heatmap

It is also interesting to observe that this SARS-COV-2 virus continues to mutate such as the Delta variant which infects many people in Indonesia. From [22], it is stated that this viral mutation of the Delta varian is occurred in Spike protein, meaning that other protein of the virus such as the 3CLPro, PLPro, and RdRp do not change. From this study, most compounds target the 3Clpro and PLPro of SARR-COV-2. With these findings, it can be said that those compounds still have the potential to inhibit the main protease protein of SARS-CoV-2 Delta variant. But of course, all of this still has to be validated by in vitro, in vivo, and clinical testing.

V. CONCLUSION

On 3CLPro, the best probabilities are compound alpha-eudesmol, cadinol, and alpha-cadinol, while on PLPro, compound with best probabilities are 1,4-naphthalenedione, germacrene, and alpha-terpinolene, on RdRp, compound with best probabilities are 1,4-naphthalenedione, alpha-cadinol, and alpha-eudesmol. Some main compounds of eucalyptus, 1,8-cineole (0.51 on 3CLPo, 0.66 on PLPro) (50.64 %), α-pinene (12.17 %), 2,4-pentanediol (11.44 %), α-terpinene (0.81 on 3CLPro, 0.9322 on PLPro, 0.86 on RdRp)(10.98 %), and limonene (3.61%) have a varied probabilities but alpha-terpinene is the most standout, with more than 80% probabilities on 3CLPro, PLPro and RdRp. Based on our study, eucalyptus can be one of the potential antiviral herbs to use as a prevention against covid-19.

ACKNOWLEDGMENT

This research is supported by Ministry of Research, Technology and Higher Education, Indonesia, under Competitive Basic Research Grant from Directorate of Higher Education, Indonesia, 2021, contract no. 2100/IT3. L1/PN/2021.

REFERENCES

[1] T. I. Oprea, J. E. Bauman, C. G. Bologa, T. Buranda, A. Chigaev, B. S. Edwards, et al., "Drug repurposing from an academic perspective," *Drug Discov Today Ther Strateg*, vol. *8*(3-4), 2011, pp. 61-69.

[2] S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, et al., "Drug repurposing: progress, challenges and recommendations," *Nat Rev Drug Discov,* vol. 18, no.1, 2019, pp. 41-58.

[3] S. L. Senanayake, "Drug repurposing strategies for COVID-19," editorial, *Future Drug. Discov*, vol. 2, no.2, 2020. editorial

[4] S. L. Kinnings, N. Liu, P. J. Tonge, R. M. Jackson, L. Xie, & P. E. Bourne, "A machine learning-based method to improve docking scoring functions and its application to drug repurposing", *Journal of chemical information and modeling*, vol. 51, no.2, 2011, pp. 408-419.

[5] J. Hu, Y. Li, J.-Y. Yang, H.-B. Shen, and D.-J. Yu, "GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure," *Computational Biology and Chemistry*, vol. 60, pp. 59–71, 2016.

[6] L. Erlina, R. I. Paramita, W. A. Kusuma, F. Fadilah, A. Tedjo, I. P Pratomo, N. S. Ramadhanti, A. K. Nasution, F. K. Surado, A. Fitriawan, et al., "Virtual Screening on Indonesian Herbal Compounds as COVID-19 Supportive Therapy: Machine Learning and Pharmacophore Modelling Approaches," preprint, BMC, 2021.

[7] L. Breiman, *Machine Learning*, vol. 36, no. 1/2, pp. 85–103, 1999.

[8] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J Photogramm Remote Sens,* vol. *114*, 2016, pp. 24-31.

[9] R. Díaz-Uriarte, and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, vol. *7*, no. 1, pp. 1-13.

[10] E. D. Coelho, J. P. Arrais, and J. L. Oliveira, "Computational discovery of putative leads for drug repositioning through drug-target interaction prediction," *PLoS computational biology*, vol. 12, no. 11, 2016 e1005219.

[11] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.

[12] How many trees in a random forest?. In *International workshop on machine learning and data mining in pattern recognition* C. Wu, Y. Liu, Y. Yang, P. Zhang, W. Zhong, Y. Wang, et al., "Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods," *Acta Pharm Sin B,* 2020;(PG-).

[13] G. Li, and E. De Clercq, "Therapeutic options for the 2019 novel coronavirus (2019-nCoV)", *Nat Rev Drug Discov,* vol. 19, 2020, pp. 149-150.

[14] S. Günther, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, et al, "SuperTarget and Matador: resources for exploring drug-target relationships," *Nucleic Acids Res,* vol. 36, 2007, pp. D919-D922.

[15] 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," *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2020.

[16] M. Bhasin and G. P. S. Raghava, "Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition," *Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23262–23266, 2004.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, et al. "Scikit-learn: Machine learning in Python," *J Mach Learn Res*, vol. 12, 2011, pp. 2825-2830.

[18] S. Madeddu, A. Marongiu, G. Sanna, C. Zannella, D. Falconieri, S. Porcedda, A. Manzin, and A. Piras, "Bovine Viral Diarrhea Virus (BVDV): A Preliminary Study on Antiviral Properties of Some Aromatic and Medicinal Plants," *Pathogens*, vol. 10, no. 4, p. 403, 2021.

[19] M. Asif, M. Saleem, M. Saadullah, H. S. Yaseen, and R. Al Zarzour, "COVID-19 and therapy with essential oils having antiviral, anti-inflammatory, and immunomodulatory properties," *Inflammopharmacology*, vol. 28, no. 5, pp. 1153–1161, 2020.

[20] D. S. Tshibangu, A. Matondo, E. M. Lengbiye, C. L. Inkoto, E. M. Ngoyi, C. N. Kabengele, G. N. Bongo, B. Z. Gbolo, J. T. Kilembe, D. T. Mwanangombo, C. M. Mbadiko, S. O. Mihigo, D. D. Tshilanda, K.-T.-N. Ngbolua, and P. T. Mpiana, "Possible Effect of Aromatic Plants and Essential Oils against COVID-19: Review of Their Antiviral Activity," *Journal of Complementary and Alternative Medical Research*, pp. 10–22, 2020.

[21] D. Planas, D. Veyer, A. Baidaliuk, I. Starpoli, F. Guivel-Benhassine, M. F. Rajah, C. Planchais, F. Porrot, N. Robillard, J. Puech, et. al. "Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization." Nature (2021). https://doi.org/10.1038/s41586-021-03777-9