Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary: Rapid Review: Open Access Journal **Vol. 6, No. 4, October 2024, pp: 478-488; eISSN: 2656-8632**

RESEARCH ARTICLE **OPEN ACCESS**

Manuscript received July 17, 2024; revised August 13, 2024; accepted Agusut 16, 2024; date of publication October 20, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/jeeemi.v6i4.506

How to cite: Daffa Nur Fiat, Syifabela Suratinoyo, Indri Claudia Kolang, Injilia Tirza Ticoalu, Nadira Tri Ardianti, Reza Michelly Cantika Mawara, Daniel Febrian Sengkey, Angelina Stevany Regina Masengi and Alwin Melkie Sambul, "Comparative Analysis of Hepatitis C virus Genotype 1a (Isolate 1) using Multiple Regression Algorithms and Fingerprinting Techniques", Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 4, pp. 478-488, October 2024.

Comparative Analysis of Hepatitis C virus Genotype 1a (Isolate 1) using Multiple Regression Algorithms and Fingerprinting Techniques

Daffa Nur Fiat[1], Syifabela Suratinoyo[1], Indri Claudia Kolang[1], Injilia Tirza Ticoalu[1], Nadira Tri Ardianti[1], Reza Michelly Cantika Mawara[1], Daniel Febrian Sengkey[1,3], Angelina Stevany Regina Masengi[2,3], and Alwin Melkie Sambul[1,3]

[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
[3]BioMolecular Laboratory, Universitas Sam Ratulangi, Polyclinic Building, 3rd Floor, Kleak, Manado 95115, North Sulawesi, Indonesia

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

ABSTRACT Approximately 70 million people worldwide have been infected with the Hepatitis C virus (HCV), which is a significant health problem associated with severe liver diseases such as acute hepatitis cirrhosis, and chronic hepatitis. HCV, which is part of the Flaviviridae family, encodes a single polyprotein consisting of 3010 amino acids, which is processed into 10 polypeptides, including the core structural protein, envelope glycoproteins E1 and E2, and nonstructural proteins such as NS3, NS5A, and NS5B. This study used machine learning to analyze HCV polyproteins with various fingerprinting methods and regression algorithms. Specifically, the default model outperformed the hyperparameter-tuned model in both R^2 and Adjusted R^2 values. Gradient Boosting Regression (GBR) and Random Forest Regression (RFR) were particularly effective, with GBR achieving the highest R^2 using PubChem fingerprints. Although hyperparameter tuning did not significantly improve model predictions, feature importance analysis provided valuable insights into specific features that impact model performance. This analysis not only simplifies the model but also highlights the most relevant features of the HCV polyprotein. These findings have practical implications for the development of more targeted and effective antiviral therapies. By identifying key features that influence model performance, this study contributes to a deeper understanding of the HCV polyprotein structure, which is critical for developing new drugs. Insights gained from feature importance analysis can drive future efforts in drug discovery, allowing for more precise therapies targeting HCV, potentially leading to improved global health outcomes.

INDEX TERMS drug discovery, hepatitis C virus, machine learning, polyprotein, regression algorithm.

I. INTRODUCTION

Hepatitis C virus (HCV) is a pathogen that chronically infects approximately 70 million people worldwide and poses a significant global health threat[1][2]. Hepatitis C (HCV) is a major cause of liver disease, with one-third of individuals with chronic HCV infection developing liver cirrhosis. In chronic HCV infection, host immune factors together with the actions of HCV proteins that enhance viral persistence and immune system dysregulation impact the immunopathogenesis of HCV-induced hepatitis[3]. Hepatitis C virus (HCV) is an RNA virus that enters the body through the mouth[4]. Transmission or blood plasma, unsafe health care, blood or blood plasma transfusion are unsafe injection practices [5]. Hepatitis C virus does not distinguish continental boundaries, so it can be found in almost all places where humans live [3]. It is estimated that although the

incidence of HCV infection appears to be decreasing in developed countries, secondary deaths associated with HCV infection will continue to increase over the next 20 years [6]. Thus, although many data suggest that HCV infection could be eliminated in the next 15-20 years with focused therapeutic strategies [7][8]. A good understanding of HCV infection should be necessary to develop strategies to prevent new infections. As a member of the Flaviviridae family, HCV can cause a variety of clinical outcomes, including acute hepatitis, chronic hepatitis, cirrhosis, or the establishment of an asymptomatic carrier state that can last a lifetime [9]. The chronic nature of HCV infection emphasizes the importance of developing better predictive models to understand the disease course and improve clinical management[10][11].

As with other members of the Flaviviridae family, the Hepatitis C virus (HCV) genome encodes a single polyprotein. This polyprotein consisting of 3010 amino acids is processed by cellular and viral proteases to produce 10 polypeptides [12]. Polypeptides range from structural proteins such as core proteins and envelope glycoproteins E1 and E2, to nonstructural proteins such as NS1, NS2, NS3, NS4A, NS4B, NS5A, and NS5B, each of which has a different role in HCV RNA replication and particle assemble[13] . Nonstructural proteins are released from the polyprotein after cleavage by HCV NS2-3 and NS3-4A proteases, while structural proteins are released by host endoplasmic reticulum (ER) signaling peptidases [14]. Polyprotein processing is initiated co-translationally and post-translationally by host and viral proteases, resulting in the formation of functional proteins of at least 10 individual structural proteins: 5′-C-E1-E2-p7-NS2-NS3-NS4A-NS4B-NS5A-NS5B-3' [13][15]. Structural proteins go into the formation of a virus particle, but nonstructural proteins contribute significantly to the replication complex and the ability of the virus to evade the host immune system [16]. For example, an NS3/4A protease complex is essential in polyprotein maturation, while NS4B has importance in the formation of the membrane mesh, a unique structure in the host cell in which viral replication occurs [16][17]. In addition, proteins such as NS5A have diverse roles, including modulation of the host immune response and involvement in the assembly of viral replication complexes [15].

A further process mediated by signal peptide peptidases also occurs at the C-terminus of capsid proteins [18]. In addition to large open reading frames encoding polyproteins, the HCV genome contains overlapping +1 reading frames that can lead to the synthesis of additional proteins [19].

New drug development often starts from an unmet clinical need. In the case of HCV it is especially important. Due to the genetic variability and drug resistance of this virus [20]. There are several effective therapies, the genetic variability of the virus and drug resistance remain major challenges. Therefore, there is an urgent need to find new therapeutic targets and develop more effective drugs.

Drug discovery programs are initiated because there are diseases or clinical conditions that have no suitable means of treatment, and this unmet clinical need is the motivation that drives the project [21][22],. The unmet clinical need becomes a driver to lead to research that generates data and develops a hypothesis using potential treatment pathways [23]. Early research, which is often the case in academia, generates data to develop hypotheses regarding the inhibition or activation of pathways that will produce a treatment effect in a disease state. The result of this research is the selection of targets that may require further validation before proceeding to the drug discovery state to gain support for drug discovery efforts [23]. During lead discovery, an in-depth search is conducted to find small molecules such as drugs or biological therapeutics, usually called development candidates, which will proceed to the preclinical testing stage, and if successful, to the clinical development stage, and ultimately become marketable drugs [24].

Machine learning is an important and rapidly growing area in computer-assisted drug discovery [25]. Machine learning algorithms are pattern recognition in understanding a mathematical relationship between empirical observations of small molecules and predicting the properties of new compounds. Compared to physical models, machine learning techniques are more efficient and scalable to large data sets without requiring large computational resources. One of the main application areas for machine learning in drug discovery is helping researchers understand and utilize the relationship between chemical structure and its biological activity or SAR [26]. Machine learning technology is built on disciplines such as statistics, mathematics, and data mining, which allows systems to learn from existing data without the need for manual reprogramming [27]. Nowadays, machine learning has become a major highlight in technology due to its capabilities and rapid development [28]. Machine learning can provide abundant data utilization, both from the internet and from data collection sensors. Machine learning algorithms such as Decision Trees, Random Forests, Support Vector Machines, and K-Nearest Neighbors have shown high accuracy in diagnosing and predicting diseases such as eye disorders, diabetes, Alzheimer's, heart disease which proved that machine learning can improve work efficiency, ensure quality, abundant utilization, and drive innovation in various industries [29][34] Machine learning is also useful in processing big data from sensors, which leads to mole cular computing [30]. however, large amounts of data can be a problem, and integrating human knowledge is crucial [31].

Although much research has been done on Hepatitis C, predictive and analytical approaches are still lacking. We can use Machine Learning to analyze Hepatitis C virus genotype 1a (isolate 1) (HCV) by using several regression algorithms and fingerprinting techniques. These techniques allow the identification of specific patterns and characteristics of polyproteins that may not be visible through conventional analysis. Therefore, this study aims to fill the gap and provide new insights that can be used in the development of more effective therapies.

FIGURE 1 shows the steps of the Hepatitis C virus (HCV) analysis process starting with data collection and screening, after which data labeling and standardization are

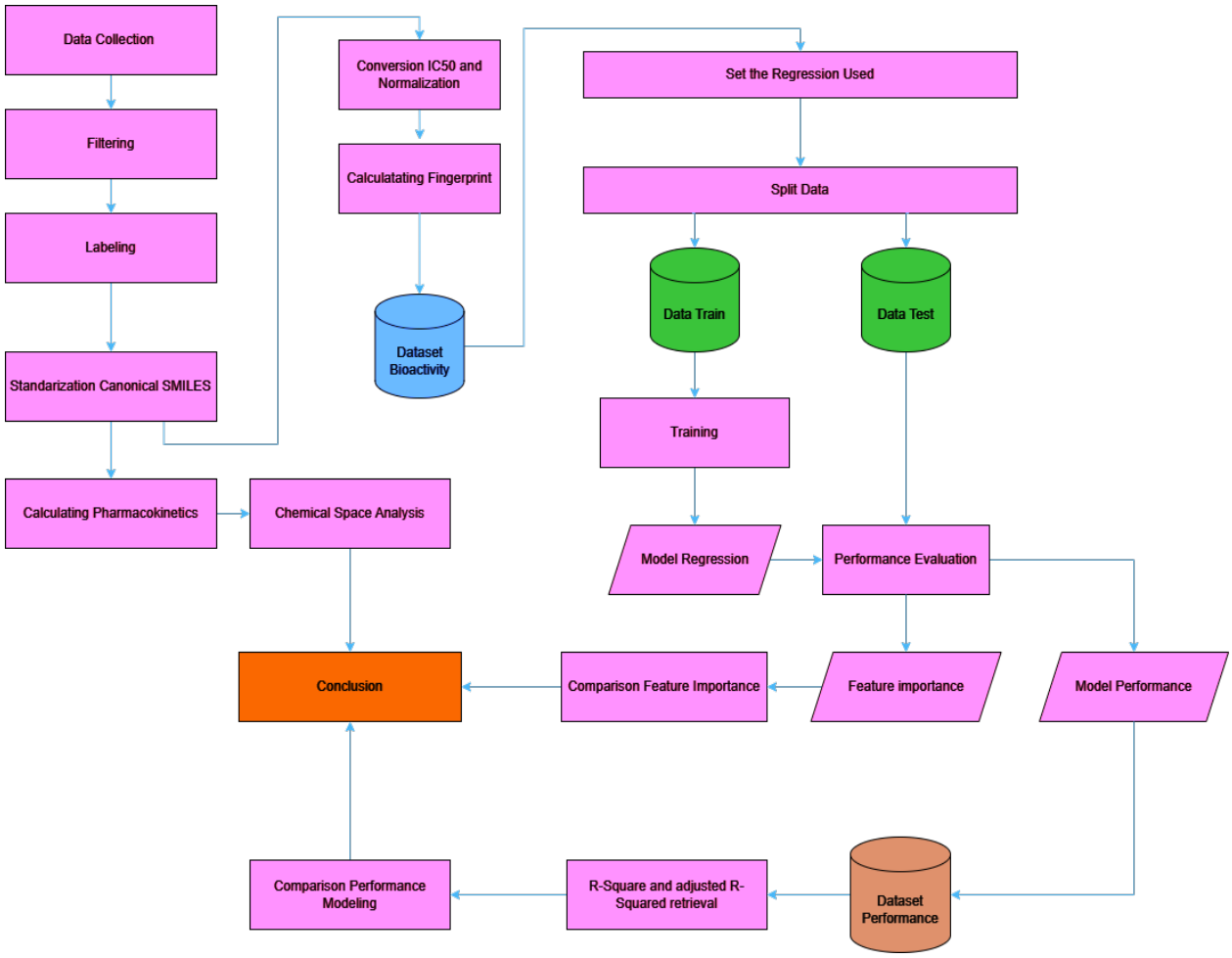

FIGURE 1. Research Course

carried out using the SMILES format, then the data is processed to calculate pharmacokinetics and obtain molecular fingerprints, as well as conversion and normalization of IC50 data. Furthermore, chemical space is carried out with the aim of understanding the relationship between molecular structure and biological activity. When the dataset is ready, it will be divided into training data and testing data to train and evaluate the regression model to be selected. Once the model has been trained, its performance will be evaluated and analyzed to determine the most influential features. Then, the results of the model's performance will be compared using several metrics, such as R-Square and Adjusted R-Square, and finally conclusions will be drawn based on the results of the analysis that has been carried out. This systematic approach is expected to provide new insights in the development of more effective therapies for HCV. Furthermore, it will discuss some machine learning regression algorithms and fingerprinting methods, this research aims to make an analysis of Hepatitis C virus (HCV) polyproteins. This study will detail the entire process, starting from data preparation, where bioactivity data derived from the ChEMBL database will be cleaned and labeled, after which pharmacokinetic and fingerprinting calculations will be performed. The data will then undergo hyperparameter tuning to optimize performance, and feature importance analysis will be performed to identify key predictors. The results will then be compared across different models to determine the most effective approach for HCV polyprotein analysis, ultimately aiding in drug discovery and therapy development.

II. METHODS

Using several machine learning regression algorithms and several fingerprint methods, this research will create a comparative Hepatitis C polyprotein virus analysis. In this section, the flow of the research process will be explained, starting from data preparation to entering the feature important section by going to the chemical space analysis section, namely drug discovery and also data comparison.

A. DATA PREPARATION

This section, we collected and prepared data to compare Hepatitis C virus polyproteins using different fingerprinting methods and predefined algorithms. The bioactivity dataset was obtained from ChEMBL through its website and an accessible Python Application Programming Interface (API) [35].

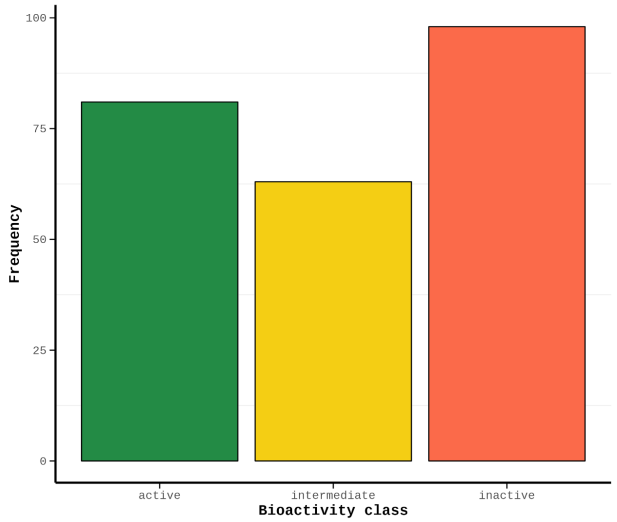

FIGURE 2. Frequency of compounds by bioactivity class.

In clinical trials, each principal investigator may need to provide consent to share data with participants. In the case of prospective studies, where study data is collected prospectively, consent is also required. After clinical approval, relevant data needs to be accessed, queried, de-identified, and securely stored [36]. The target retrieved was CHEMBL4620. At that time, there were 243 data on the bioactivity of inhibitors against HCV replication polyprotein stored in the CHEMBL database. Before data processing is carried out, data cleaning will be carried out to improve the quality of more accurate data and avoid anomalous data, with data cleaning ensuring data is ready to enter the analysis stage [37]. In the next stage, a compound labeling will be carried out which is divided into 3 namely 'active', 'between', and 'inactive', this labeling will display the frequency of each bioactivity class which can be seen in FIGURE 2. Labeling it with a value less than 1000nM will be declared that the compound is 'active'. If the value of the compound is between 1000nM and 10000nM then the compound will be declared 'between', and if it is above 10000nM then it will be declared an 'inactive' compound [35].

Pharmacokinetics is a pharmaceutical science about a drug that can be absorbed by the body. After performing a labeling, a pharmacokinetic calculation is performed including 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). By performing a pharmacokinetic calculation, an overview of the statistical analysis can be made. Then, the value of the Inhibitory Concentration 50% (IC50) is changed to the negative logarithm of the IC50 value (pIC50), then a fingerprint is formed to continue modeling using a regression algorithm.

B. FEATURES EXTRACTION

Fingerprinting in chemistry involves creating unique representations of chemical structures to facilitate the analysis of molecular similarities and properties. PubChem fingerprints are created through a series of pre-designed molecular substructures. These fingerprints are binary vectors that indicate the presence or absence of these substructures in the molecule. PubChem fingerprinting is particularly useful in database searching and similarity assessment as it offers a representation of complex molecular features. This method tends to perform well in detecting compounds with similar biological activities due to its detailed and complete substructure coding [38].

MACCS keys are a series of 166 predefined structures that provide a binary representation of a molecule. Each key corresponds to the presence (1) or absence (0) of a particular structural feature. MACCS keys are widely used due to their ease and efficiency in structural similarity searches. These keys present a balance between computational efficiency and the ability to retrieve appropriate chemical features, making them suitable for a wide range of chemical applications [39].

E-State fingerprinting provides information about the electronic state and topological status of atoms in a molecule. This fingerprint considers both the innate electronic properties of atoms and their interrelationships within the molecule. This dual consideration allows E-State fingerprinting to capture more diverse aspects of molecular structure, which can be very beneficial in quantitative structure-activity relationship (QSAR) studies and other applications that require detailed electronic information [40].

C. HYPERPARAMETER TUNING MODELING

In this study there are two stages of modeling to get an evaluation of the performance of the regression model. First, modeling with default parameters that have been determined from the algorithm used and modeling with hyperparameter tuning by determining the parameters you want to use and also the feature selection you want to use which aims to get the optimal hyperparameter combination. The purpose of this approach is to compare R^2 and Adjusted R^2 from both models.

D. FEATURE IMPORTANCE

Feature Importance is a step to measure the influence of each input feature on machine learning model prediction. By knowing which features affect the prediction results the most, practitioners and researchers can get information about the data and models used. Feature importance analysis helps in:

1. MODEL SIMPLIFICATION

Identifying and removing less important features can simplify the model, making it easier to interpret and faster to execute. This process can improve model efficiency and also help reduce complexity [41].

2. DATA UNDERSTANDING

Knowing which features are most influential helps researchers understand the factors underlying predictive outcomes. This is particularly useful in fields like

biomedicine, where a deep understanding of important factors can guide clinical interventions.

3. MODEL PERFORMANCE IMPROVEMENT

Focusing on the most important features can improve model performance by reducing noise from irrelevant features.

4. MODEL VALIDATION

Knowing the important features allows further validation of the model results with domain expertise to ensure that the results obtained make logical and scientific sense.

In this study, the three best models out of 12 tested models will be selected based on the highest R^2 value obtained from hyperparameter tuning testing. The algorithms used to perform hyperparameters are Gradient Boosting Regression (GBR), Random Forest Regression (RFR), AdaBoost Regression (ABR) and Hist Gradient Boosting Regression (HistGBR). It is important to analyze only the 3 best models to understand the contribution of each molecular feature to prediction.

III. RESULT

The research results of the previously described methods will be presented as a result of the hyperparameter tuning that has been carried out for two different modelings using several selected regression and fingerprint algorithms. These results include data analysis.

A. CHEMICAL SPACE ANALYSIS

Chemical Space analysis is shown to evaluate the distribution of molecular descriptors in the Hepatitis c virus polyprotein dataset. In this Chemical Space, there are four descriptors analyzed LogP as in FIGURE 3(a), MW as in FIGURE 3(b), NumAcceptors as in FIGURE 3(d), and NumHDonors as in FIGURE 3(c). In evaluating the distribution of these descriptors, there are significant differences between groups in the dataset, here we use the Kruskal-Wallis test, the analysis results are made into one table.

Based on the results of the Kruskal-Wallis test, it was concluded that there were significant differences in only three descriptors: LogP, MW, NumHAcceptors. This result shows that the p value is very small ($p < 0.05$), which makes it possible to reject the null hypothesis (H0).

1. LOGP

The values in TABLE 1 show that the LogP distribution is significantly different among the descriptor groups.

2. MW

The MW values in TABLE 1 are also significantly different between groups. These significant differences indicate variations in molecular size among the descriptor groups.

3. NUMHACCEPTORS

The number of hydrogen acceptors is significantly different between groups. The number of hydrogen acceptors is an important parameter in the ability of compounds to interact through hydrogen bonding.

TABLE 1
Kruskal-Wallis test results between bioactivity classes according to descriptors

Descriptor	Statistics	p	alpha	Interpretation
LogP	26.4150	0.000002	0.05	Diffrent distribution (reject H0)
MW	45.2368	1.502970e-10	0.05	Diffrent distribution (reject H0)
NumHDonors	0.054	0.973327	0.05	Diffrent distribution (reject H0)
NumHAcceptors	51.8459	5.518266e-12	0.05	Same distribution (fail to reject H0)

However, NumHDonors differs from the other descriptors in that it does not show that there is a significant difference in the distribution of the number of hydrogen donors between groups. This suggests that there are differences in each of the main chemical descriptors.

B. REGRESSION MODEL PERFORMANCE

Machine learning techniques in performing correlation calculations using Gradient Boosting Regression (GBR), Random Forest Regression (RFR), AdaBoost Regression (ABR) and Hist Gradient Boosting Regression (HistGBR) regression models [42]. In this model, default modeling will be performed with the dataset divided into 80% training and 20% testing to evaluate and compare the four ensemble regression models based on several performance evaluation metrics R^2 and Adjusted R^2. The default model is implemented using the library available from 'skit-learn', in this model the parameters are not specified manually.

TABLE 2
Parameter used for modelling.

Algorithm	Parameters	Value
HistGBR	variance_threshold__threshold	4
	regressor__max_iter	3
	regressor__max_depth	3
	regressor__min_samples_leaf	4
	regressor__learning_rate	2
ABR	variance_threshold__threshold	4
	regressor__n_estimators	5
	regressor__learning_rate	4
	regressor__loss	3
GBR	variance_threshold__threshold	4
	regressor__max_depth	3
	regressor__min_samples_leaf	4
	regressor__max_features	3
	regressor__learning_rate	3
RFR	variance_threshold__threshold	4
	regressor__n_estimators	3
	regressor__max_samples	2
	regressor__min_samples_leaf	2
	regressor__max_depth	3

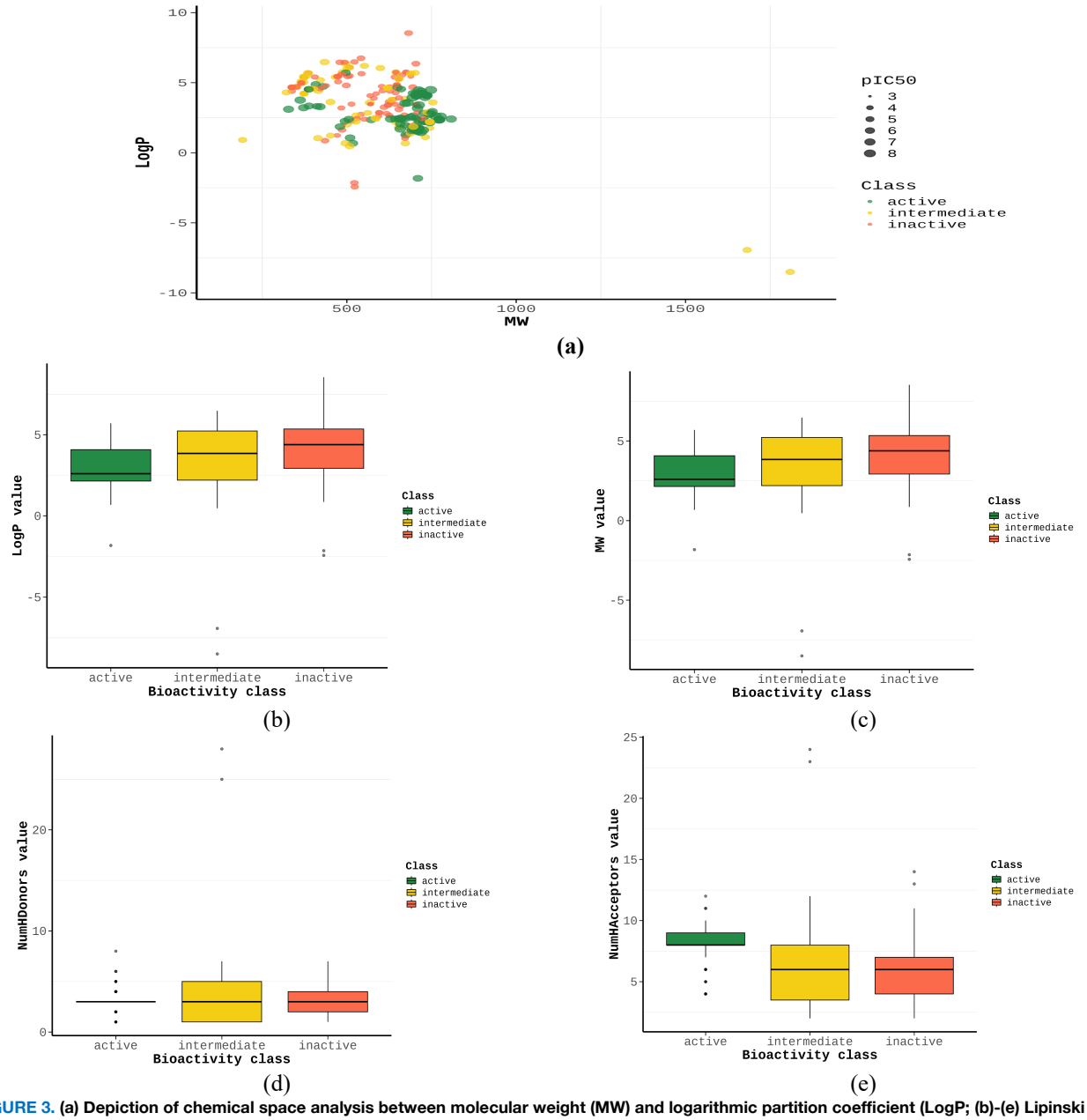

FIGURE 3. (a) Depiction of chemical space analysis between molecular weight (MW) and logarithmic partition coefficient (LogP; (b)-(e) Lipinski descriptor distribution

After that, a re-modeling using hyperparameter tuning is performed to obtain the optimal combination of parameters. Hyperparameter tuning is done to improve the performance of the model during training.

TABLE 3
Performance of algorithmic regression using the default model

Model	RMSE	R^2	Adj R^2	Fingerprint
GBR	0.527205	0.844771	1.008945	PubChem
RFR	0.582567	0.810458	1.076454	MACCS
GBR	0.598568	0.799904	1.080711	MACCS
HistGBR	0.611532	0.791142	1.084245	MACCS
RFR	0.582567	0.774218	1.013010	PubChem
GBR	0.665663	0.752531	1.383178	Estate
ABR	0.680017	0.741743	1.104171	MACCS
HistGBR	0.740795	0.681264	1.401349	Estate
HistGBR	0.611532	0.740717	1.014941	PubChem
ABR	0.680017	0.736641	1.015176	PubChem
RFR	0.731285	0.701334	1.462451	Estate
ABR	0.748709	0.686932	1.484750	Estate

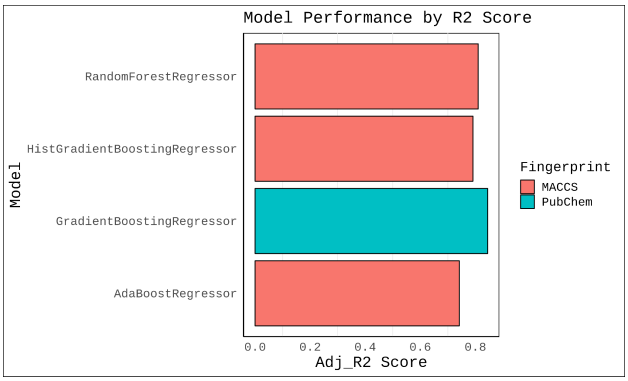

FIGURE 4. Algorithmic regression performance using the default 4 best models (R² shown).

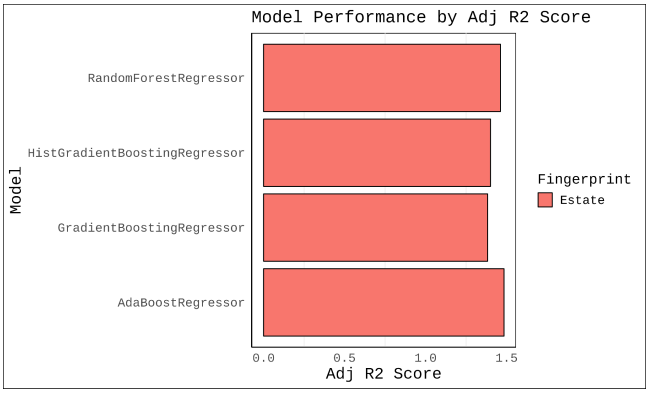

FIGURE 5. Algorithmic regression performance using the default 4 best models (Adjust R² displayed).

TABLE 2 summarizes the parameters used in the hyperparameter tuning modeling process for the four algorithms used in machine learning: Histogram-based Gradient Boosting Regression (HistGBR), AdaBoost Regression (ABR), Gradient Boosting Regression (GBR), and Random Forest Regression (RFR). This table shows the specific parameters adjusted in optimizing model performance such as maximum iterations, tree depth, learning rate, and number of estimators, to optimize model performance. In addition, the value of each parameter varies.

TABLE 3 shows a regression model performance using the default model with R² evaluation metric. Evaluated models GBR, RFR, ABR, and HistGBR, the results show that Gradient Boost Regression (GBR) using PubChem fingerprint is the best among other algorithms and fingerprints. However, when viewed in the evaluation metric of the Adjusted R² section. The results provide information that the Gradient Boost Regression (GBR) shows very good performance among other models, but the difference is in the fingerprint where the best performance is using the Estate fingerprint and the biggest difference is the Adjusted R² adjustment on the Estate fingerprint, this shows a higher Adjusted R² value compared to the MACCS and PubChem fingerprints.

From the analysis using the default model, a modeling is then performed using hyperparameter tuning. Where the function of hyperparameter tuning is to regulate the optimization of model performance by adjusting what parameters will be used when modeling, where the parameters are parameters that are not learned by the model.

TABLE 4
Performance of algorithmic regression using hyperparameter tuning model

Model	RMSE	R²	Adj R²	Fingerprint
GBR	0.5697	0.8187	10.104	PubChem
HistGBR	0.6117	0.7910	10.843	MACCS
HistGBR	0.6134	0.7899	10.121	PubChem
GBR	0.6168	0.7875	10.857	MACCS
RFR	0.6465	0.7665	10.135	PubChem
RFR	0.6511	0.7633	10.955	MACCS
HistGBR	0.6730	0.7470	13.917	Estate
ABR	0.6743	0.7460	11.024	MACCS
GBR	0.6746	0.7459	13.935	Estate
ABR	0.6787	0.7428	10.148	PubChem
RFR	0.6968	0.7288	14.199	Estate
ABR	0.7367	0.6969	14.693	Estate

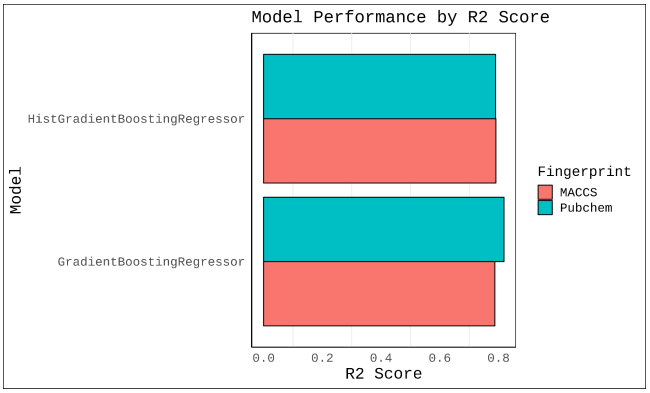

FIGURE 6. Algorithmic regression performance using hyperparameter tuning 4 best models (R² shown).

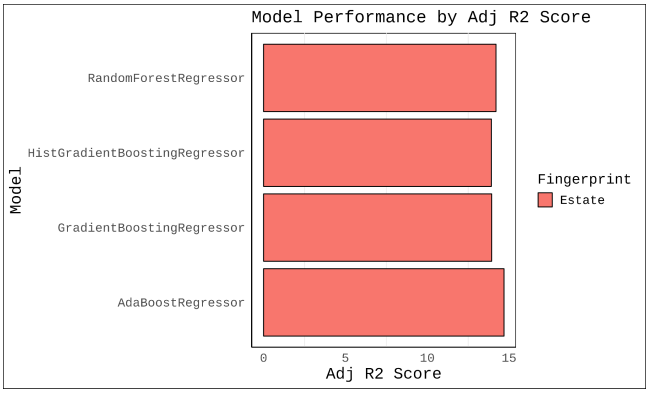

FIGURE 7. Algorithmic regression performance using hyperparameter tuning 4 best models (Adjust R² shown).

It can be seen in TABLE 4 If you look at the comparison between TABLE 4 and TABLE 3 there are differences where there are several models that have increased when doing hyperparameter tuning and there are also models that have

decreased as well and when viewed from FIGURE 6 it can be seen that the best models in hyperparameter tuning are HistGBR and GBR where both are best using MACCS and Pubchem fingerprints. In FIGURE 7 we can see that the best Adjust R^2 4 models are in all algorithms but if we compare TABLE 3 with TABLE 4 there is an increase in Adjust R^2 in all algorithms using the Estate fingerprint.

C. FEATURE IMPORTANCE

In performing regression analysis using machine learning models such as Gradient Boosting Regression (GBR), Random Forest Regression (RFR), AdaBoost Regression (ABR), and Hist Gradient Boosting Regression (HistGBR), it is very important in analyzing regression algorithms to be able to understand which features have a better influence on the target. Important features not only improve a model's accuracy, but can provide a good understanding to be able to explore more important components in the model.

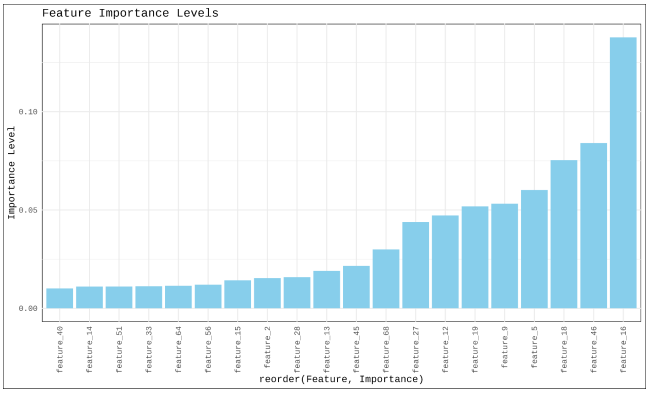

FIGURE 8. Feature importance of the model using MACCS fingerprint and Random Forest Regression algorithm.

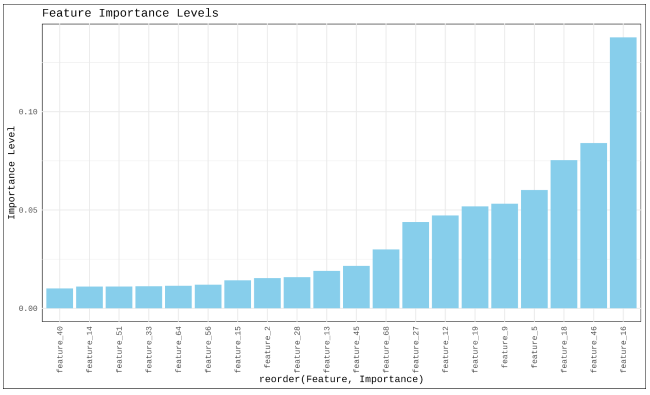

FIGURE 9. Feature importance of the model using MACCS fingerprint and Gradient Boost Regression algorithm.

Evaluating the contribution to model performance is part of the feature importance selection process. Evaluation metrics such as adjusted R^2 and Adjusted R^2 can be used to evaluate whether the model is well adapted to the complexity of the data. Hyperparameter tuning is also used to optimize model parameters that are not learned during training to achieve the best performance. FIGURE (8-10) shows how important this feature is in performing a machine learning modeling. The horizontal and vertical axes show the level of feature importance to the model performance. The higher bar is the most important or most influential feature in modeling and the lower bar is the least influential feature in modeling.

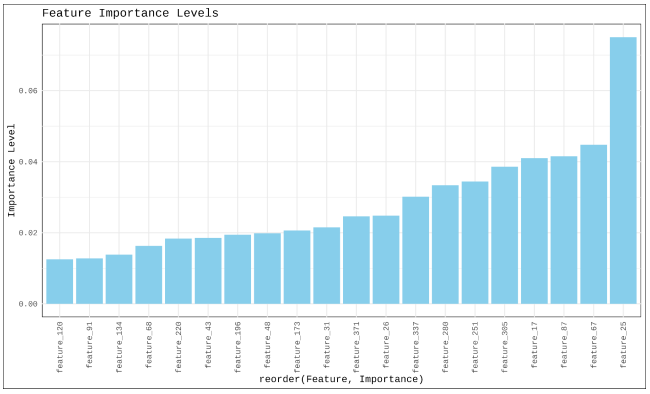

FIGURE 10. Feature importance of the model using Pubchem fingerprint and Gradient Boost Regression algorithm.

The evaluation of feature importance can help to understand the role of features used when performing a hyperparameter tuning. This feature can allow us to reduce features or dimensions that can be reduced to reduce model complexity and improve model prediction.

IV. DISCUSSION

A better understanding of the structure and function of HCV polypeptides may pave the way for more effective and safe therapies for Hepatitis C. Polypeptides are proteins composed of amino acids and are an important component of the HCV virus. They play an important role in viral replication, disease pathogenesis, and response to therapy. Finding optimal algorithms and fingerprinting techniques for Hepatitis C Virus (HCV) polypeptide analysis is an important step in developing more accurate, efficient, and informative analysis methods [32]. This study will evaluate the performance of various algorithms for HCV polypeptide analysis. Algorithms will be evaluated based on their accuracy, sensitivity, specificity, and speed. This research will compare various fingerprint techniques for HCV polypeptide analysis. Fingerprint techniques will be evaluated based on their ability to produce unique and informative protein representations. Based on the evaluation results, this study will determine the most effective fingerprint algorithms and techniques for HCV polypeptide analysis [33]. Through a comprehensive evaluation of various fingerprinting algorithms and techniques, our research shows that in the Regression algorithm using the Default Model which shows that Gradient Boost Regression (GBR) using Pubchem fingerprint is the best based on R^2 results, but when viewed from Adjust R^2 the best performance is in the model using Gradient Boost Regression (GBR) algorithm with E-State fingerprint, both of which produce optimal performance for data results in HCV polypeptide analysis.

Research on HCV polypeptide analysis has the potential to generate valuable information that can contribute significantly to a better understanding of the virus, improve diagnosis and monitoring methods, and pave the way for the development of more sophisticated and accurate therapies. The information obtained from fingerprint analysis can also be used to design new drugs that target HCV viral proteins more specifically and effectively. The results of this study are expected to make a significant contribution to the fight against Hepatitis C and improve public health globally.

Recent studies have shown that HCV Genotype 1a provides better virological response to antiviral therapy compared to Genotype 1b. This finding was important for our study, which used regression analysis to evaluate factors affecting therapy outcomes in HCV Genotype 1a [43]. This comparison allowed us to assess how characteristics of HCV genotype 1a influence response to therapy in a broader context. Although this article focuses on differences in virologic response, the analysis methodology used, may differ from the regression approach applied in our study. Differences in analysis techniques may affect the interpretation of the results and reveal factors that may not have been identified in this study. In addition, findings regarding differences in virologic response may provide additional context to the results of our regression analysis, which aimed to identify variability in therapy response based on HCV genotype. Thus, our study is expected to add to the understanding of antiviral therapy effectiveness and HCV genotype variability in clinical medicine.

However, some limitations of this study should be noted. For example, this study may be affected by the quality and quantity of the data used, as well as possible biases in the selection of features and algorithms. Further research is needed to overcome these limitations and to explore how the same techniques can be applied to larger and more diverse datasets.

V. CONCLUSION

This study aims to find a model with the optimal regression algorithm and fingerprint technique to analyze the polyprotein of Hepatitis C Virus (HCV). With the model carried out using two modeling techniques where the first uses default modeling, the second uses a hyperparameter tuning model to be able to compare models and analyze which model is good to use. Chemical space analysis using the Kruskal Wallis test to be able to see significant differences in three of the four molecular descriptors (LogP, MW, and NumHAcceptors) across the dataset, showing a varied distribution among bioactivity classes, while NumHDonors did not show significant differences. Modeling using the default model shows that the best model is GBR as the algorithm and its fingerprint uses Pubchem seen from the highest R^2, while when viewed from Adjust R^2 the best model is GBR with its fingerprint is Estate. After using the hyperparameter tuning model, HistGBR and GBR show good performance especially on MACSS and Pubchem fingerprints, with an increase in Adjust R^2 on all algorithm lags showing Estate fingerprints. The analysis highlights that understanding the selection of important features is very influential in understanding the complexity of the model and making the model more accurate. This study can help provide insight into drug selection in hepatitis C polyprotein virus using machine learning modeling by looking at R^2 and Adjusted R^2. This study provides insights into the selection of optimal regression algorithms and fingerprinting techniques to analyze Hepatitis C Virus polyproteins which are expected for future research to be focused on expanding the dataset, exploring additional molecular descriptors, and integrating using more advanced machine learning techniques to further improve the accuracy and application of the model in drug discovery efforts for Hepatitis C.

REFERENCES

[1] A. D. Dearborn and J. Marcotrigiano, "Hepatitis C Virus Structure: Defined by What It Is Not," *Cold Spring Harb Perspect Med*, vol. 10, no. 1, p. a036822, Jan. 2020, doi: 10.1101/cshperspect.a036822.

[2] J. Dubuisson, "Hepatitis C virus proteins," *World J Gastroenterol*, vol. 13, no. 17, p. 2406, 2007, doi: 10.3748/wjg.v13.i17.2406.

[3] D. Chigbu, R. Loonawat, M. Sehgal, D. Patel, and P. Jain, "Hepatitis C Virus Infection: Host–Virus Interaction and Mechanisms of Viral Persistence," *Cells*, vol. 8, no. 4, p. 376, Apr. 2019, doi: 10.3390/cells8040376.

[4] A. Petruzziello, S. Marigliano, G. Loquercio, A. Cozzolino, and C. Cacciapuoti, "Global epidemiology of hepatitis C virus infection: An up-date of the distribution and circulation of hepatitis C virus genotypes," *World J Gastroenterol*, vol. 22, no. 34, p. 7824, 2016, doi: 10.3748/wjg.v22.i34.7824.

[5] M. B. Butt *et al.*, "Diagnosing the Stage of Hepatitis C Using Machine Learning," *J Healthc Eng*, vol. 2021, pp. 1–8, Dec. 2021, doi: 10.1155/2021/8062410.

[6] H. Razavi *et al.*, "Chronic hepatitis C virus (HCV) disease burden and cost in the United States," *Hepatology*, vol. 57, no. 6, pp. 2164–2170, Jun. 2013, doi: 10.1002/hep.26218.

[7] H. Wedemeyer *et al.*, "Strategies to manage hepatitis <scp>C</scp> virus (<scp>HCV</scp>) disease burden," *J Viral Hepat*, vol. 21, no. s1, pp. 60–89, May 2014, doi: 10.1111/jvh.12249.

[8] N. K. Martin, M. Hickman, S. J. Hutchinson, D. J. Goldberg, and P. Vickerman, "Combination Interventions to Prevent HCV Transmission Among People Who Inject Drugs: Modeling the Impact of Antiviral Treatment, Needle and Syringe Programs, and Opiate Substitution Therapy," *Clinical Infectious Diseases*, vol. 57, no. suppl_2, pp. S39–S45, Aug. 2013, doi: 10.1093/cid/cit296.

[9] A. Grakoui, C. Wychowski, C. Lin, S. M. Feinstone, and C. M. Rice, "Expression and identification of hepatitis C virus polyprotein cleavage products," *J Virol*, vol. 67, no. 3, pp. 1385–1395, Mar. 1993, doi: 10.1128/jvi.67.3.1385-1395.1993.

[10] M. A. Konerman, Y. Zhang, J. Zhu, P. D. R. Higgins, A. S. F. Lok, and A. K. Waljee, "Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data," *Hepatology*, vol. 61, no. 6, pp. 1832–1841, Jun. 2015, doi: 10.1002/hep.27750.

[11] M. A. Konerman, S. Yapali, and A. S. Lok, "Systematic review: identifying patients with chronic hepatitis C in need of early treatment and intensive monitoring – predictors and predictive models of disease progression," *Aliment Pharmacol Ther*, vol. 40, no. 8, pp. 863–879, Oct. 2014, doi: 10.1111/apt.12921.

[12] F. Penin, J. Dubuisson, F. A. Rey, D. Moradpour, and J.-M. Pawlotsky, "Structural biology of hepatitis C virus," *Hepatology*, vol. 39, no. 1, pp. 5–19, Jan. 2004, doi: 10.1002/hep.20032.

[13] D. Pascut, M. Hoang, N. N. Q. Nguyen, M. Y. Pratama, and C. Tiribelli, "HCV Proteins Modulate the Host Cell miRNA Expression Contributing to Hepatitis C Pathogenesis and Hepatocellular Carcinoma Development," *Cancers (Basel)*, vol. 13, no. 10, p. 2485, May 2021, doi: 10.3390/cancers13102485.

[14] K. E. Reed and C. M. Rice, "Overview of Hepatitis C Virus Genome Structure, Polyprotein Processing, and Protein Properties," 2000, pp. 55–84. doi: 10.1007/978-3-642-59605-6_4.

[15] Y. Zhang, X. Zhao, J. Zou, Z. Yuan, and Z. Yi, "Dual role of the amphipathic helix of hepatitis C virus NS5A in the viral polyprotein

cleavage and replicase assembly," *Virology*, vol. 535, pp. 283–296, Sep. 2019, doi: 10.1016/j.virol.2019.07.017.
[16] S. Barik, "Suppression of Innate Immunity by the Hepatitis C Virus (HCV): Revisiting the Specificity of Host–Virus Interactive Pathways," *Int J Mol Sci*, vol. 24, no. 22, p. 16100, Nov. 2023, doi: 10.3390/ijms242216100.
[17] R. Khandia, A. A. Khan, N. Karuvantevida, P. Gurjar, I. V. Rzhepakovsky, and I. Legaz, "Insights into Synonymous Codon Usage Bias in Hepatitis C Virus and Its Adaptation to Hosts," *Pathogens*, vol. 12, no. 2, p. 325, Feb. 2023, doi: 10.3390/pathogens12020325.
[18] J. McLauchlan, "Intramembrane proteolysis promotes trafficking of hepatitis C virus core protein to lipid droplets," *EMBO J*, vol. 21, no. 15, pp. 3980–3988, Aug. 2002, doi: 10.1093/emboj/cdf414.
[19] A. D. Branch, D. D. Stump, J. A. Gutierrez, F. Eng, and J. L. Walewski, "The Hepatitis C Virus Alternate Reading Frame (ARF) and Its Family of Novel Products: The Alternate Reading Frame Protein/F-Protein, the Double-Frameshift Protein, and Others," *Semin Liver Dis*, vol. 25, no. 01, pp. 105–117, Feb. 2005, doi: 10.1055/s-2005-864786.
[20] K. Lin, "Development of novel antiviral therapies for hepatitis C virus," *Virol Sin*, vol. 25, no. 4, pp. 246–266, Aug. 2010, doi: 10.1007/s12250-010-3140-2.
[21] C. Granchi, "Biological Activity of Natural and Synthetic Compounds," *Molecules*, vol. 27, no. 12, p. 3652, Jun. 2022, doi: 10.3390/molecules27123652.
[22] A. S. Verkman, "Drug discovery in academia," *American Journal of Physiology-Cell Physiology*, vol. 286, no. 3, pp. C465–C474, Mar. 2004, doi: 10.1152/ajpcell.00397.2003.
[23] A. Roy, "Early Probe and Drug Discovery in Academia: A Minireview," *High Throughput*, vol. 7, no. 1, p. 4, Feb. 2018, doi: 10.3390/ht7010004.
[24] J. Hughes, S. Rees, S. Kalindjian, and K. Philpott, "Principles of early drug discovery," *Br J Pharmacol*, vol. 162, no. 6, pp. 1239–1249, Mar. 2011, doi: 10.1111/j.1476-5381.2010.01127.x.
[25] A. Varnek and I. Baskin, "Machine Learning Methods for Property Prediction in Chemoinformatics: *Quo Vadis* ?," *J Chem Inf Model*, vol. 52, no. 6, pp. 1413–1437, Jun. 2012, doi: 10.1021/ci200409x.
[26] S. M. Ali, M. Z. Hoemann, , Jeffrey Aubé, G. I. Georg, L. A. Mitscher, and L. R. Jayasinghe, "Butitaxel Analogues: Synthesis and Structure−Activity Relationships," *J Med Chem*, vol. 40, no. 2, pp. 236–241, Jan. 1997, doi: 10.1021/jm960505t.
[27] A. Raj, "A Review on Machine Learning Algorithms," *Int J Res Appl Sci Eng Technol*, vol. 7, no. 6, pp. 792–796, Jun. 2019, doi: 10.22214/ijraset.2019.6138.
[28] R. R. Reddy, C. Mamatha, and R. G. Reddy, "A Review on Machine Learning Trends, Application and Challenges in Internet of Things," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, Sep. 2018, pp. 2389–2397. doi: 10.1109/ICACCI.2018.8554800.
[29] P. K. Donepudi, "Automation and Machine Learning in Transforming the Financial Industry," *Asian Business Review*, vol. 9, no. 3, pp. 129–138, 2019, doi: 10.18034/abr.v9i3.494.
[30] C. Zhao *et al.*, "Multiscale Construction of Bifunctional Electrocatalysts for Long-Lifespan Rechargeable Zinc–Air Batteries," *Adv Funct Mater*, vol. 30, no. 36, Sep. 2020, doi: 10.1002/adfm.202003619.
[31] C. Deng, X. Ji, C. Rainey, J. Zhang, and W. Lu, "Integrating Machine Learning with Human Knowledge," *iScience*, vol. 23, no. 11, p. 101656, Nov. 2020, doi: 10.1016/j.isci.2020.101656.
[32] D. Moradpour and F. Penin, "Hepatitis C Virus Proteins: From Structure to Function," 2013, pp. 113–142. doi: 10.1007/978-3-642-27340-7_5.
[33] M. Golizeh *et al.*, "Proteomic fingerprinting in HIV/HCV co-infection reveals serum biomarkers for the diagnosis of fibrosis staging," *PLoS One*, vol. 13, no. 4, p. e0195148, Apr. 2018, doi: 10.1371/journal.pone.0195148.
[34] S. Tang, "Applications of Machine Learning in the Industry of Healthcare," *Highlights in Science, Engineering and Technology*, vol. 1, pp. 87–96, Jun. 2022, doi: 10.54097/hset.v1i.432.
[35] D. F. Sengkey and A. Masengi, "Regression Algorithms in Predicting the SARS-CoV-2 Replicase Polyprotein 1ab Inhibitor: A Comparative Study," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 1, pp. 1–10, Dec. 2023, doi: 10.35882/jeeemi.v6i1.338.
[36] M. J. Willemink *et al.*, "Preparing Medical Imaging Data for Machine Learning," *Radiology*, vol. 295, no. 1, pp. 4–15, Apr. 2020, doi: 10.1148/radiol.2020192224.
[37] F. Ridzuan and W. M. N. Wan Zainon, "A Review on Data Cleansing Methods for Big Data," *Procedia Comput Sci*, vol. 161, pp. 731–738, 2019, doi: 10.1016/j.procs.2019.11.177.
[38] H. Kuwahara and X. Gao, "Analysis of the effects of related fingerprints on molecular similarity using an eigenvalue entropy approach," *J Cheminform*, vol. 13, no. 1, p. 27, Dec. 2021, doi: 10.1186/s13321-021-00506-2.
[39] E. Fernández-de Gortari, C. R. García-Jacas, K. Martinez-Mayorga, and J. L. Medina-Franco, "Database fingerprint (DFP): an approach to represent molecular databases," *J Cheminform*, vol. 9, no. 1, p. 9, Dec. 2017, doi: 10.1186/s13321-017-0195-1.
[40] D. Boldini, D. Ballabio, V. Consonni, R. Todeschini, F. Grisoni, and S. A. Sieber, "Effectiveness of molecular fingerprints for exploring the chemical space of natural products," *J Cheminform*, vol. 16, no. 1, p. 35, Mar. 2024, doi: 10.1186/s13321-024-00830-3.
[41] M. Saarela and S. Jauhiainen, "Comparison of feature importance measures as explanations for classification models," *SN Appl Sci*, vol. 3, no. 2, p. 272, Feb. 2021, doi: 10.1007/s42452-021-04148-9.
[42] S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara, and M. Roja Edinburgh, "Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, Jan. 2021, pp. 989–995. doi: 10.1109/Confluence51648.2021.9377137.
[43] A. M. Pellicelli *et al.*, "HCV genotype 1a shows a better virological response to antiviral therapy than HCV genotype 1b," *BMC Gastroenterol*, vol. 12, no. 1, p. 162, Dec. 2012, doi: 10.1186/1471-230X-12-162.

AUTHOR BIOGRAPHY

DAFFA NUR FIAT is a student of Informatics Engineering at Sam Ratulangi University, Manado, North Sulawesi, he is the first of two children, born in Sidoarjo on April 8, 2004 and he comes from South Tangerang. he is a person who has a passion for doing new things and can be trusted in working in a team. He has participated in the TECHOVERSE game project in 2023 in the unity engineer section which is a big project made by the UNITY Organization and he has won 1st place in the satria data competition in the Statistical Infographic Competition division and 1st place in the gemastik competition in the game development division at the unsrat level in 2024. now he is a member of UNITY as a mentor of the Unity Engineer division and now he is participating in a big UNITY project called SPARK in the machine learning engineer division. He also now serves as the chairman of the reasoning UKM.

SYIFABELA SURATINOYO is an active 5th semester student at the Department of Informatics Engineering, Faculty of Engineering, Sam Ratulangi University (UNSRAT), Manado, North Sulawesi. Born on July 20, 2004, in Gorontalo, Indonesia, she shows a strong commitment to her studies and extracurricular activities. Active in various organizations and communities both on and off campus, she is dedicated to continuous learning and personal growth. Her academic focus includes a deep interest in design, specifically UI/UX, where she utilizes her skills in HTML, CSS, as well as the design application Figma to create engaging and user-friendly digital interfaces. She also has valuable experience in team-based competitions. Her diverse involvement highlights her broad capabilities and potential as a professional in informatics and design.

INDRI CLAUDIA KOLANG is an active student of class 2022 at Sam Ratulangi University, Faculty of Engineering, Department of Electrical Engineering, Informatics Engineering study program. She is the first daughter of 2 children and was born in Palu, June 25, 2004 and studied at SMA Negeri 1 Kauditan. During her college years, she has developed several projects and has done some research with teams and individuals. He has an interest in the field of design related to his major and has an interest in several other fields. He is also active in activities and organizations on and off campus. He is also active in ministry in his church and in the area where he lives and is active in the student service organization in the Faculty of Engineering. Currently, he will focus more on developing soft skills and hard skills.

INJILIA TIRZA TICOALU is an active 4th semester student at Informatics Engineering Study Program, Electrical Engineering Department, Faculty of Engineering, Sam Ratulangi University Manado. Besides pursuing her academic education, she has a great interest in leadership and community service. Currently, she serves as the Deputy Director of Human Resources Department at Maleosan.ID, where she focuses on team development and internal system management. In addition, she also plays an active role as a Member of the Multimedia Documentation & Publication Division in the Tatelu Region Exemplary Youth Association for the period 2022-2024, contributing fully to strengthening the organization's media presence. He is committed to making a positive impact in the field of technology as well as society.

NADIRA TRI ARDIANTI is an active 4th semester student at Sam Ratulangi University, Manado, Faculty of Engineering, Informatics Engineering Study Program. born on March 29, 2005 in Manado. He studied at SMA Negeri 8 Manado. He comes from a harmonious and religious family and has two older brothers. She has experience in project teams where she worked on project assignments during college. He is also active in organizations or activities on and off campus. He has an interest in Front End and an interest in web design, especially UI/UX. In between his busy schedule on campus, he also pursues his hobby of running and participates in running events. His hope for the future is to graduate on time and work in a reputable company by becoming a good Front End Developer and making his loved ones happy.

REZA M. C. MAWARA is a student of Informatics Engineering Study Program, Department of Electrical Engineering, Faculty of Engineering, Sam Ratulangi University, Manado, Indonesia. He studied at Don Bosco Lembean Catholic High School, majoring in Natural Sciences. Born in Lembean on February 1, 2004, he is the first of four children. She has experience in project teams where she worked on various project tasks during college. He has an interest in the field of Design and Artificial Intelligence, especially in the field of UI/UX, he is also active in various committee activities and student organizations on campus. Her technical skills include programming in HTML and CSS, as well as experience in using software development tools such as Figma.

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.

Alwin Melkie Sambul earned his undergraduate degree in Electrical Engineering from Universitas Sam Ratulangi in 2003. His Master's and Ph.D. degrees in Biomedical Engineering, from Kumamoto University, Japan, are completed in 2011 and 2015, respectively. Dr. Sambul is currently the Head of the Electrical Engineering Department at the Faculty of Engineering, and also part of the Bioinformatics team at the Biomolecular Laboratory. Both offices he holds are within the Universitas Sam Ratulangi. His research interest is in Biomedical Engineering, especially in brainwave signaling.