Journal of Physics:
Conference Series

PAPER • **OPEN ACCESS**

COVID-19 mRNA Vaccine Degradation Prediction Using LR and LGBM Algorithms

To cite this article: Soon Hwai Ing *et al* 2021 *J. Phys.: Conf. Ser.* **1997** 012005

COVID-19 mRNA Vaccine Degradation Prediction Using LR and LGBM Algorithms

Soon Hwai Ing[1], Azian Azamimi Abdullah[1, 2], Nor Hazlyna Harun[3] and Shigehiko Kanaya[4]

[1]Faculty of Electronic Engineering Technology, Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia.
[2]Medical Devices and Life Sciences Cluster, Sport Engineering Research Centre, Centre of Excellence (SERC), Universiti Malaysia Perlis (UniMAP), Perlis, Malaysia.
[3]Data Science Research Lab, School of Computing, School of Computing, Universiti Utara Malaysia Sintok, Kedah.
[4]Computational Systems Biology Lab, Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma, Nara, Japan.

E-mail: azamimi@unimap.edu.my

Abstract. The threatening Coronavirus which was assigned as the global pandemic concussed not only the public health but society, economy and every walks of life. Some measurements are taken to stifle the spread and one of the best ways is to carry out some precautions to prevent the contagion of SARS-CoV-2 virus to uninfected populaces. Injecting prevention vaccines is one of the precaution steps under the grandiose blueprint. Among all vaccines, it is found that mRNA vaccine which shows no side effect with marvellous effectiveness is the most preferable candidates to be considered. However, degradation had become its biggest drawback to be implemented. Hereby, this study is held with desideratum to develop prediction models specifically to predict the degradation rate of mRNA vaccine for COVID-19. Two machine learning algorithms, which are, Linear Regression (LR) and Light Gradient Boosting Machine (LGBM) are proposed for models development using Python language. Dataset comprises of thousands of RNA molecules that holds degradation rates at each position from Eterna platform is extracted, pre-processed and encoded with label encoding before loaded into algorithms. The results show that LGBM (0.2447) performs better than LR (0.3957) for this study when evaluated with the RMSE metric.

1. Introduction

SARS-CoV-2, an air-borne virus that resulted millions of death since the end of the year 2019 and treatment specifically to the cause of this virus, the COVID-19 is yet to be excavated [1]. Even effective vaccines for preventive are currently under study. Although mRNA vaccines are the one that entrusted with highest hopes amongst, it has a drawback of rapid degradation.

Refer to research done by Wadhwa et al. [2] in year 2020, upon the occurrence of in vitro transcription, degradation will greatly reduce the yields of mRNA. In addition, this paper claims that under refrigerated cool chain transport condition, the half-life of the mRNA vaccine may have a half-life of 900 days with a rate of at least 2% of degradation every 30 days [2]. It is worth noting that with a digression of temperature to around 37°C, or merely 2 units drift of pKa value, it is believed to reduce dramatically the half-life of a vaccine to 5 days and 10 days, respectively [2]. In addition to this, with

the presence of Mg^{2+} with a temperature of 37°C, customarily the condition for in vitro transcription, the vaccine half-life will further reduce to not more than 2 hours [2]. The result remains the same even after reducing the concentration Mg^{2+}, pH value nor temperature to alleviate the hydrolysis where mRNA vaccine is unstable to as degradation still occurs during the transcription process [2]. After vaccination into a human body, a 5-days half-life [2] is estimated for the mRNA vaccine.

Although Abbasi stated that researchers believe that this drawback can be solved with a second dose-regimen of the candidate vaccine [3], but this degradation issue should not be overlooked as the potency of a vaccine can never be restored nor regained once it is lost. Parenthetically, owing to a finale high-stability mRNA vaccine is yet to be developed, and it may take up months to years for a positive result to be bear, what we can do as for now is to implement the candidate vaccine in hand to control this pandemic. And with this, the stability of the vaccine should be as precise as possible, which in other words of saying, the degradation rate of the vaccine that easily alters by both extrinsic and intrinsic factors should be clear and definite, and of course, same goes to the future-successfully-developed vaccine as well.

As stated, at this juncture, the study on the degradation of mRNA vaccine is extremely crucial. Nevertheless, study and research on predicting the degradation of mRNA or even vaccines are extremely limited, not to mention regarding mRNA vaccines for COVID-19. The only research accessible currently on this topic is a study published by Ankit Singhal in late 2020 with LSTM, GRU and GCN algorithms, evaluated with RMSE showing GCN-based model (0.249) is the finest [4].

Thence, this study is focused to develop models and design rules using machine learning algorithms to predict the degradation rate of mRNA vaccine. The proposed model will be used to predict the degradation rates at each base of an RNA molecule which was trained on a subset of an Eterna dataset that comprises of 6034 RNA molecules that hold degradation rates at each position.

2. Methodology

The main purpose of this study is to develop reliable model that able to predict the degradation rate of COVID-19 mRNA vaccine. In general, there are 3 main stages, namely, data pre-processing, model training and performance evaluating.

For this study, the BPPs NumPy file that holds the probability for each base of RNA to be paired is extracted together with the train and test datasets from Eterna database platform [5] that consists of numbers of RNA molecules that hold degradation rates at each position. A number of features will be engineered from the BPPs dataset and only those features that are suitable for prediction among all those engineered features will be selected.

Pre-processing will then take part to eliminate noises and to organize the data for training and testing purpose. After done processing the data, converting non-numerical data to numerical data, 2 algorithms, the LR and LGBM are proposed and trained with train dataset for models development. The performance of the model develop on test dataset is evaluated with Root Mean Square Error, RMSE. From the RMSE values resulted, we will choose the best model amongst. The general methodology flow chart is shown in Figure 1.

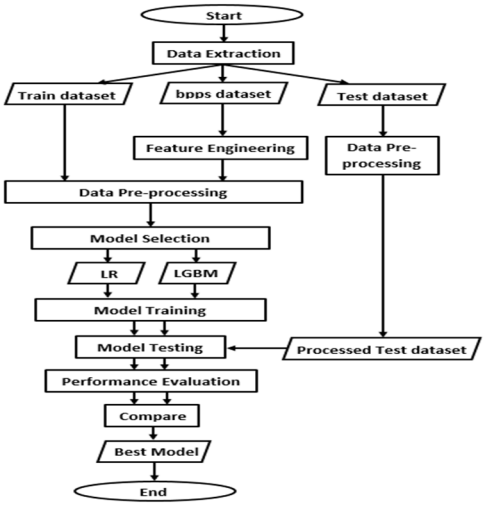

Figure 1. The methodology flow chart.

2.1. Dataset

Dataset is a constellation of data, may be in form of an array, a data structure whose elements are denoted from the same data type, or in a database table form that mayhap possess different data types. The three most common data types in ML are numerical, categorical and ordinal data. In the dataset, the rows of the data are designated as instances also known as the observation collected, while the columns could represent either the features or the classes. Features are the independent characteristics, while classes are the dependent output that we intend to predict. The dataset can be in form of strings, dates or a more labyrinthine type too. Datasets usually assorted into 2 different sets, training dataset and testing dataset, each with different purposes.

2.1.1. Train Dataset. The train dataset of this study encompassed of 2400 instances with 19 features, including *index* and *id*. It is fortunate that there are no missing values in the dataset for each instances and features. Table 1 shows the list of features for this study.

Table 1. List of features.

Features	Description
index	Numerical order list for each sample
id	Identifier for each sample
sequence	A combination of A, U, G, and C bases for each sample that depict the RNA sequence. Either 1 × 107 or 1 × 130 characters long.
structure	An alignment of (,), and . characters that illustrate the pairing state of the RNA. '(' and ')'indicates paired base while '.', unpaired interaction. Length corresponds to *'sequence'* feature.
predicted_loop_type	Describe the structure of each character in *'sequence'* that delineate RNA structures with code named 'bpRNA'. B: Bulge; E: dangling End; H: Hairpin loop; I: Internal loop; M: Multiloop; S: paired "Stem"; X: eXternal loop
signal_to_noise	Determine the quality of the sample. Higher SNR, higher quality.
SN_filter	Denoted with 1 if sample passed both the 2 filter conditions, else 0. 2 conditions, considering only the first 68 bases of RNA samples sequence in Train dataset: (1) Minimum value > − 0.5 across all 5 classes. (2) Mean SNR > 1.0 across all 5 classes.
seq_length	Depict the length of *'sequence'* of RNA samples. Either 107 or 130.

seq_scored	68 for *'seq_length'* with 107 and 91 for 130's. Depicting the number of positions used in scoring with predicted values, analogous to the length of all 5 classes together with their 'error'.
reactivity_error	An array of calculated floating point error akin to 'reactivity' column obtained in experiment juxtapose to the physical world situation.
deg_error_Mg_pH10	An array of calculated floating point error akin to 'deg_Mg_pH10' column obtained in experiment juxtapose to the physical world situation.
deg_error_pH10	An array of calculated floating point error akin to 'deg_pH10' column obtained in experiment juxtapose to the physical world situation.
deg_error_Mg_50C	An array of calculated floating point error akin to 'deg_Mg_50C' column obtained in experiment juxtapose to the physical world situation.
deg_error_50C	An array of calculated floating point error akin to 'deg_50C' column obtained in experiment juxtapose to the physical world situation.
reactivity	An array of floating point numbers, either 1 × 68 or 1 × 91 vectors, which convey the probability of the base to be paired.
deg_Mg_pH10	An array of floating point numbers, either 1 × 68 or 1 × 91 vectors, which convey the degradation or fragility of the linkage in each base under high pH condition (pH10), with the presence of Mg.
deg_pH10	An array of floating point numbers, either 1 × 68 or 1 × 91 vectors, which convey the degradation or fragility of the linkage in each base under high pH condition (pH10), without the presence of Mg.
deg_Mg_50C	An array of floating point numbers, either 1 × 68 or 1 × 91 vectors, which convey the degradation or fragility of the linkage in each base under high temperature condition (50°C), with the presence of Mg.
deg_50C	An array of floating point numbers, either 1 × 68 or 1 × 91 vectors, which convey the degradation or fragility of the linkage in each base under high temperature condition (50°C), without the presence of Mg.

2.1.2. Test Dataset. For this study, the training dataset embodied by over 3000 instances, to be precise, 3634 instances with 7 features. To be noticed, the classes of this study, a total of 5, are the reactivity of each base and degradations under variegated environmental factors, pH and temperature, with the presence or absence of magnesium ions. Figure 2 below shows the first 5 instances of test dataset.

index	id	sequence	structure	predicted_loop_type	seq_length	seq_scored
0	id_00073f8be	GGAAAAGUACGACUUGAGUACGGAAAACGUACCAACUCGAUUAAAA...	((((((((((.(((((.....))))))))) (((((((.....	EEEEEESSSSSSSSSSBSSSSSHHHHHSSSSSSSSSSSSSSSHHH...	107	68
1	id_000ae4237	GGAAACGGGUUCCGCGGAUUGCUGCUAAUAAGAGUAAUCUCUAAAU...	.....(((((..(((((((...(((((..... (((((....)))))....	EEEEESSSSIISSSSSSIIISSSSSIIIIISSSSHHHHSSSSIIII...	130	91
2	id_00131c573	GGAAAACAAAACGGCCUGGAAGACGAAGGAAUUCGGCGCGAAGGCC...	..........((.(((.(.(..((..((. (((...)))).)))...	EEEEEEEEEEESSISSSISISIISSIISSIISSSSHHHSSSSIISS...	107	68
3	id_00181fd34	GGAAAGGAUCUCUAUCGAAGGAUAGAGAUCGCUCGCGACGGCACGA...	.....((((((((((....))))))))))(((((..((. (((.....	EEEEEESSSSSSSSSSHHHHSSSSSSSSSSSSSSSIISSISSSHH...	107	68
4	id_0020473f7	GGAAACCCGCCCGCGCCCGCCCGCGCUGCUGCCGUGCCUCCUCUCC...	..... ((((((((((((((((((((((((((((((((((....	EEEEESSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS...	130	91

Figure 2. First 5 rows instances with features of test dataset.

2.1.3. Bpps Dataset. Bpps is the abbreviation for base-pairing probabilities, educe that bpps symmetric square matrix NumPy file that escorts both train and test datasets possesses probability of forming a base pair for each base of the RNA. The bpps attached are the NumPy arrays calculated with algorithms developed by [5]. The bpps matrices are prepared for all the instances in both the train and test datasets, one for each row, each base in the sequence each.

2.2. Data Pre-processing

Data Pre-processing is the percussive manoeuvre before analyzing the data with a view to transmogrify the raw noisy data to a clean yet simpler data to minimize disparage of quality of data analysis as erroneous, inadequate or inconsequential data will entail faulty prediction that give rise to dire performance [6].

2.2.1. Handling Missing Data. Inadequacy of information or missing data is an inevitable challenge that customarily occurs in genuine data sources while analysing data. The occurrence of this type of mistake may arise from the account of lost or oblivion [6]. Inapplicability of features to instances so does the nonchalant feature value are some additional reasons behind the deficiency of feature values [6]. Missing value issue should be solved as most of the algorithms do not have the authority to endorse missing values in dataset fed. With Python command '*.isnull()*', there is no missing value in datasets extracted.

2.2.2. Data Cleaning. It is worth notifying that mRNA with noisy results ought not to be used for actual vaccine development. Therefore, to ensure only sublime samples are fed to the model, the instances were filtered based on stipulated criteria referring to the corresponded signal_to_noise and SN_filter features as mentioned in Table 1. As mentioned, if all the criteria are passed by the instances, SN_filter will be denoted with 1. Thence, to ensure the models are able to perform as pre-eminence as possible, train dataset is strained, with only those instances that passed the SN_filter are considered, resulting in reduction of number of instances in train dataset from 2400 to 1589.

2.2.3. Label Encoding. In ML, data can be categorised into 3 main explicit data type categories: numerical, categorical and ordinal. Although some of the models can handle diverse types of data, there are still a considerable amount of algorithms cannot meet the desired aptitude. Consequently, data are recommended to be modified from non-numerical datatype data to numerical datatype data for proper processing. Thence, label encoding which is a simple yet splendid encoding technique with impressive performance [7] is proposed to encode the 3 non-numerical features, *sequence*, *structure and predicted_loop_type.* The arrays, either 1 × 107 or 1 × 130 characters long, will first be split into 1 × 1 character long, resulting an increase in the number of instances from thousands to hundred thousand, together with their corresponded features, the 5 classes and the error columns. After done producing 1 × 1 variable instances, the characters is label-encoded individually as shown in Table 2.

Table 2. Label Encoding.

RNA sequence		Structure		Predicted Loop Type	
Sequence	Index	Sequence	Index	Sequence	Index
A	0	.	0	B	0
U	1	(	1	E	1
G	2	)	2	H	2
C	3			I	3
				M	4
				S	5
				X	6

After the label encoding is done on the filtered train dataset and test dataset, the instances increase dramatically from 1589 to 108052 for train dataset and from 3634 samples to 457953 samples for the test dataset.

2.3. Feature Engineering

The preeminent feature engineering [8] is a process of preliminarily processing the raw data into a more presentable and breezier form features that are compatible with algorithms for modelling to improve their prediction performances [9]. Superb quality features are believed to suffice in mitigating the storage load, saving the storage space and cutting down effectively the processing time required [10]. In this prediction study, we will manoeuvre feature engineering on bpps matrices dataset into features in form of aggregate functions, max, mean, nznbr, std, and sum which will then be analyzed with data visualization techniques.

2.4. ML Algorithms

2 ML regression algorithms, LR and LGBM are implemented to develop models for this regression-based supervised learning study. Models' prediction errors are evaluated with RMSE.

2.4.1. Linear regression, LR. LR is a common yet well-known supervised machine learning algorithms among both rookies and experts in the field of data science [11]. LR algorithm operation is perfectly simple and understandable, leaving no doubts to the user by fitting a regression line to the data and dexterously expounding the relation between variables (dependent with independent). LR is a traditional algorithm that is well-known to have a good performance in regression problem resulting it to be proposed in this regression-based study. The off-the-shelf example can be seen from the research done by Bayrak and Ogul [12] in bioinformatics filed, predicting the true value of gene expression.

2.4.2. Light Gradient Boosting Unit, LGBM. LGBM is an enhanced gradient boosting framework that utilized decision tree (DT), tree-based learning algorithms, and therefore it is also assigned as the histogram-based DT algorithm [13]. LGBM is proved to have a better performance in speed as compared to DT [14]. At times, it may show an outshine accuracy and precision than DT too. Research done by Zhan et al. [13] proved that LGBM is reliable in bioinformatics regression prediction. On a different note, the studies that compare the performance of linear regression with LGBM is in minority, thereupon, the result of this study is worth looking forward to.

3. Result and Discussion

With the intention to determine suitable algorithms in developing prediction models for COVID-19 mRNA vaccine degradation, algorithms proposed need to be trained with and evaluated with a suitable and reliable performance metric. On the other hand, as mentioned in Section 2.3, data visualization technique will be utilized to determine suitable bpps aggregate-functions-features induced as the quality of the features will candidly dictate the virtuosity of the analyte after loaded into a model.

3.1. Data Visualization

As sequence length varies between instances in test dataset, the instances are categorized into 2 different categories, namely public test with seq_length valued at 107, and private test with seq_length equals to 130. To be noticed, only the first 68 (for seq_length = 107) and 91 (for seq_length = 130) with experimental data is considered owing to experimental constraints [5].

Hazardous features may have sprouted attributable to the incongruity in sequence length of instances in dataset, and these perilous features should be avoided from using since we can never assure that it may or may not induce the occurrence of overfitting and some other undesired problems.

To determine whether the developed bpps aggregate-functions-features induced from the bpps numpy files are suitable and safe to be utilized, distribution curves are delineated. Treacherous features will show different distribution compared to others. To handle, we may choose either to normalize them fastidiously with extreme care if we adjudge to consider them as input features to train the model, or just simply neglect them.

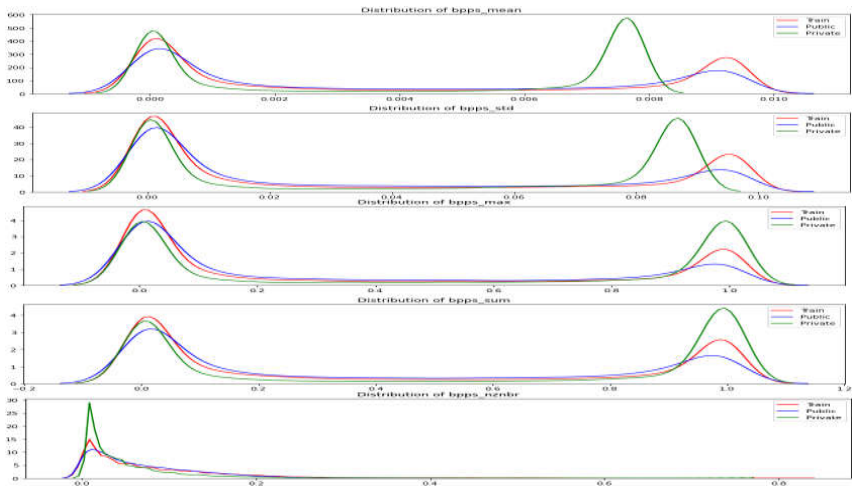

Figure 3. Bpps aggregate-functions distribution curve.

From the graphs obtained as shown in Figure 3, it seems that only both bpps_max and bpps_sum are innocuous to be utilized but on the other hand, bpps_nznbr, bpps_mean and bpps_std distribution curves show discordant between the train and private test data and therefore excluded.

3.2. RMSE Performance Metrics

RMSE is a performance metric for regression which measures the average magnitude of errors by considering the square root of the mean of the squared differences between the predictions and the ground truth [15]. The formula for RMSE metric is shown as below in equation (1) where n represents the number of instances.

RMSE is a negative-oriented scoring technique which implies the lower the RMSE value, the better the performance of the model. The score is ranged from zero to positive infinity due to the presence of square applied to the difference between ground truth and predicted values. In addition, RMSE has a magnificent performance when coping with large error values as the difference will be more divulging after the squaring in RMSE, making RMSE more perceptive to outliers.

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(predicted_i - actual_i)^2} \quad (1)$$

3.3. Prediction Performance

To reduce the computational time but remaining the accuracy [16], the filtered train dataset is cross validated into 10 folds. The ML models' performance evaluated with RMSE on each of the classes together with their overall result are presented in Table 3.

Table 3. Prediction error for LR and LGBM resulted after evaluation with RMSE performance metric.

Models	LR	LGBM
reactivity	0.3488528039592882	0.2221544618882953
deg_Mg_pH10	0.4543870507864414	0.28265498168944153
deg_pH10	0.4495909736449766	0.260638185741347
deg_Mg_50C	0.39605440852383433	0.2376603551842918
deg_50C	0.32984539723903067	0.22019122382878797
Total	0.3957461268307142	0.24465984166643273

From Table 3, the RMSE prediction metric-evaluated performance of LR and LGBM models on the datasets shows that LGBM's prediction performance surpassed the performance of LR in this study for the given dataset across all the classes. The prediction error of LGBM shows a value of over 0.15 lower than LR. Therefore, we may deduced that LGBM is more suitable than LR for this degradation performance task assigned.

4. Conclusion

Referring to the result obtained, we may conclude that for this study, LGBM is more congruous than LR for mRNA degradation predictions. LGBM-based model's performance even surpass the models developed by [4] with LSTM, GRU and GCN each with k = 4.

Although the result is laudable, the evident constraints of this study should not be overlooked. One of the biggest defects is the length of the RNA sequence studied. In practice, mRNA vaccine for COVID-19 used to be in the range of 3000 to 4000 bases long [4] but this study wields only 107 to 130 bases. Amelioration such as acquainting the model with longer mRNA vaccine sequence and evaluate the performance error to permit better utilization and reliability in predictions.

Besides, we may consider utilized 10-fold cross-validation on the DL algorithms such as GRU proposed by [4] in future research and observe the performance whether there is any improvement in predictions then compare the result between the DL algorithms and ML algorithms proposed.

Acknowledgements
This work was partly funded by Universiti Malaysia Perlis (UniMAP).

References

[1] Zhu N *et al* 2020 A Novel Coronavirus from Patients with Pneumonia in China *N. Engl. J. Med.* **382** 8 pp 727–733
[2] Wadhwa A Aljabbari A Lokras A Foged C and Thakur A 2020 Opportunities and Challenges in the Delivery of mRNA-Based Vaccines *Pharmaceutics* **12** 2 p 102
[3] Abbasi J 2020 COVID-19 and mRNA Vaccines—First Large Test for a New Approach *JAMA* **324** 12 p 1125
[4] Singhal A 2020 Application and Comparison of Deep Learning Methods in the Prediction of RNA Sequence Degradation and Stability
[5] Wayment-Steele H K *et al* 2020 Theoretical basis for stabilizing messenger RNA through secondary structure design *bioRxiv* 262931
[6] Bhagat V Robins B and Pallavi M O 2019 Sparx - Data Preprocessing Module in *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)* 1–6
[7] Jackson E and Agrawal R 2019 Performance Evaluation of Different Feature Encoding Schemes on Cybersecurity Logs in *2019 SoutheastCon* 1–9
[8] Sun Y and Yang G 2019 Feature Engineering for Search Advertising Recognition in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* 1859–1864
[9] Oyamada M 2019 Extracting Feature Engineering Knowledge from Data Science Notebooks in *2019 IEEE International Conference on Big Data (Big Data)* 6172–6173
[10] Punmiya R and Choe S 2019 Energy Theft Detection Using Gradient Boosting Theft Detector With Feature Engineering-Based Preprocessing *IEEE Trans. Smart Grid* **10** (2) 2326–2329
[11] Sravani B and Bala M M 2020 Prediction of Student Performance Using Linear Regression in *2020 International Conference for Emerging Technology (INCET)* 1–5
[12] Bayrak T and Ogul H 2018 Data Integration for gene expression prediction in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)* 1–6
[13] Zhan Z-H You Z-H Li L-P Zhou Y and Yi H-C 2018 Accurate Prediction of ncRNA-Protein Interactions From the Integration of Sequence and Evolutionary Information *Front. Genet.* **9**
[14] Mitchell R and Frank E 2017 Accelerating the XGBoost algorithm using GPU computing *PeerJ Comput. Sci.* **3** e127
[15] Neill S P and Hashemi M R 2018 Ocean Modelling for Resource Characterization *Fundamentals of Ocean Renewable Energy* 193–235
[16] Kurniabudi S D Darmawijoyo M Y Bamhdi A M and Budiarto R 2020 CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection *IEEE Access* **8** 132911–132921.