Development of COVID-19 mRNA Vaccine Degradation Prediction System

Soon Hwai Ing
Faculty of Electronic Engineering Technology
Universiti Malaysia Perlis
Kampus Alam UniMAP Pauh Putra
Arau, Perlis, Malaysia
hwaiingsoon@gmail.com

Azian Azamimi Abdullah
Faculty of Electronic Engineering Technology
Universiti Malaysia Perlis
Kampus Alam UniMAP Pauh Putra
Arau, Perlis, Malaysia
azamimi@unimap.edu.my

Shigehiko Kanaya
Graduate School of Information Science
Nara Institute of Science and Technology (NAIST)
Ikoma, Nara, Japan
skanaya@gtc.naist.jp

2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT) | 978-1-6654-4032-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/3ICT53449.2021.9582052

***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 marvelous 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. 3 machine learning algorithms, which are, Linear Regression (LR), Light Gradient Boosting Machine (LGBM) and Random Forest (RF) are proposed for 12 models development. 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 the LGBM-based model which is trained along with auxiliary bpps features and encoded with method 1 label encoding performs the best (RMSE = 0.24466), followed by the same criteria LGBM-based model but encoded with label encoding method 2, with a difference in 0.00003 in tow the topnotch model. The RF-based model with applaudable performance (RMSE = 0.25302) even without the ubieties of the riddled bpps features in contradistinction to the training and encoding criteria of the superb mellowed LGBM-based model is worth being further cultivated for the prediction study on COVID-19 mRNA vaccines' degradation rate.**

Keywords*— *MRNA vaccine; COVID-19; Degradation; Label encoding, Machine learning algorithms; Models

I. 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. 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. [1] 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. Moreover, this paper argues 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. In addition to this, it is worth noting that 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 [1]. This paper also testifies that the result remains the same even after reducing the concentration Mg^{2+}, pH value or temperature to alleviate the hydrolysis which mRNA vaccine is unstable to as degradation still occurs during the transcription process. Furthermore, after vaccination into a human body, a 5-days half-life [1] is estimated for the mRNA vaccine.

Although the degradation drawback of the candidate vaccine may be solved with a second dose regimen, 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 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 saying, the degradation rate of the vaccine that easily alters by both extrinsic and intrinsic factors should be clear and definite, and of course, the same goes to the future-successfully-developed vaccine.

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 GRU-based model (0.263) is the finest [2]. 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.

II. 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. Dataset for this study was extracted from licensed "OpenVaccine-solves" under Eterna data-collection browser-based platform (https://eternagame.org/about/software). Eterna platform [3] is constituted by Stanford University located at Californa allied with Carnegie Mellon University to unitize brains from all around the world to solve or design RNA-related problems through gaming, puzzles to benefit biomedical-related researches. Data extracted from Eterna

978-1-6654-4032-5/21/$31.00 ©2021 IEEE

comprises of train, test and bpps datasets of RNA molecule structure with the bases' degradation rates for COVID-19 mRNA vaccines. 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 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, 3 algorithms, the LGBM, LR and RF 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.

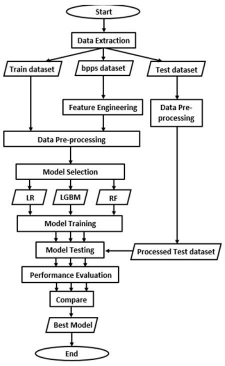


Fig. 1. The methodology flow chart.

A. Dataset

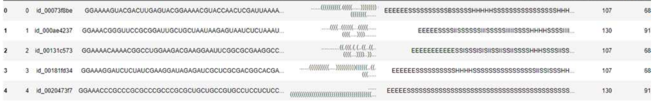
Dataset is a constellation of data, may be in form of an array, a data structure whose elements are denoted from the same data type, or in a database table form that mayhap possess different data types. The three most common data types in ML are numerical, categorical and ordinal data. In the dataset, 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.

1) Train Dataset: The train dataset of this study encompassed of 2400 instances with 19 features, namely the 'sequence', 'structure', 'predicted_loop_type', 'SN_filter', 'signal_to_noise', 'seq_length', 'seq_scored', 'reactivity_error', 'deg_error_pH10', 'deg_error_50C', 'deg_error_Mg_pH10', 'deg_error_Mg_50C', 'reactivity', 'deg_pH10', 'deg_50C', 'deg_Mg_pH10', 'deg_Mg_50C', including 'index' and 'id'. It is fortunate that there are no missing values in the dataset for each instances and features.

TABLE I. 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) Test Dataset: For this study, the test dataset embodied 3634 instances with 7 features ('index', 'id', 'sequence', 'structure', 'predicted_loop_type', 'seq_length', 'seq_scored'). 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 the test dataset.



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

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

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 [3]. 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.

B. 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.

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. Inapplicability of features to instances so does the nonchalant feature value are some additional reasons behind the deficiency of feature values. 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) 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.

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 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 are label-encoded individually with 2 different methods as shown in Table 2.

TABLE II. LABEL ENCODING

Method 1								
RNA sequence	**Char**	A	U	G	C			
	Index	0	1	2	3			
Structure	**Char**	.	(	)				
	Index	0	1	2				
Predicted Loop Type	**Char**	B	E	H	I	M	S	X
	Index	0	1	2	3	4	5	6
Method 2								
RNA sequence	**Char**	A	U	G	C			
	Index	0	1	2	3			
Structure	**Char**	.	(	)				
	Index	4	5	6				
Predicted Loop Type	**Char**	B	E	H	I	M	S	X
	Index	7	8	9	10	11	12	13

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

C. Feature Engineering

The preeminent feature engineering 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. Superb quality features are believed to suffice in mitigating the storage load, saving the storage space and cutting down effectively the processing time required. 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.

D. ML Algorithms

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

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. 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.

2) Light Gradient Boosting Machine, 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. LGBM is proposed for the reason that countless studies have proved LGBM used to have a better performance in speed as compared to DT and at times, it may show an outshine accuracy and precision than DT too. 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) Random Forest, RF: RF is an algorithm that categories under tree-based learning algorithms that alliances multitudinous DT, which could be classification or regression. RF manipulates the strength of DT and redresses its foibles of overfitting, making it a notch above DT. Some researchers with their studies have proven that RF is a splendid algorithm for bioinformatics studies, especially for classification-related analysis. With an eye to showcasing the virtue of Random Forest in regression-related problems, RF is proposed, together with LSTM and LR, for this degradation prediction study.

III. 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 and evaluated with a suitable and reliable performance metric. On the other hand, 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.

A. 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 [3]. 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.

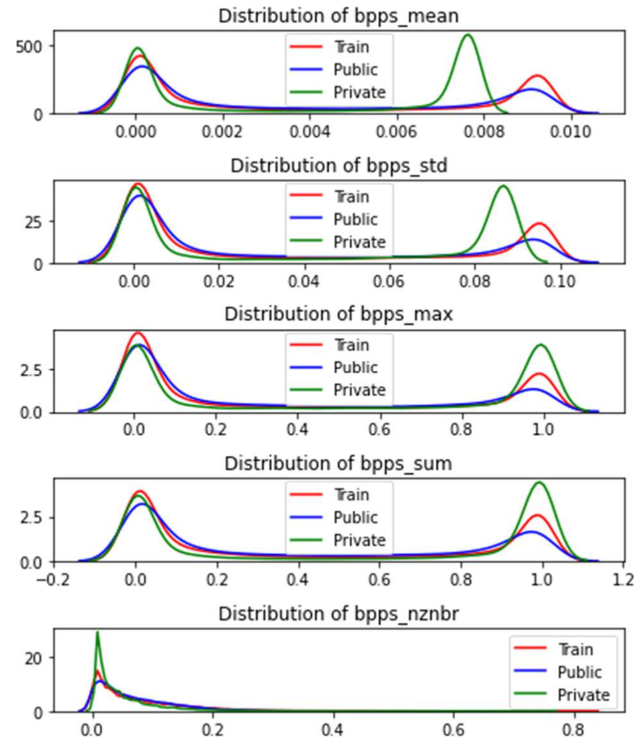


Fig. 3. Bpps aggregate-functions-features distribution curve.

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.

B. 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. The formula for RMSE metric is shown as below in equation (1) where n represents the number of instances.

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

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.

C. Prediction Performance

To reduce the computational time but remaining the accuracy, 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 III. Prediction error for LR resulted after evaluation with RMSE performance metric.

Classes	Criteria		LR
reactivity	**Label Encoding Method 1**	**With Bpps**	0.34885
		Without Bpps	0.17981
	Label Encoding Method 2	**With Bpps**	0.34075
		Without Bpps	0.17981
deg_Mg_pH10	**Label Encoding Method 1**	**With Bpps**	0.45439
		Without Bpps	0.36638
	Label Encoding Method 2	**With Bpps**	0.44210
		Without Bpps	0.36638
deg_pH10	**Label Encoding Method 1**	**With Bpps**	0.44959
		Without Bpps	0.52981
	Label Encoding Method 2	**With Bpps**	0.40772
		Without Bpps	0.52981
deg_Mg_50C	**Label Encoding Method 1**	**With Bpps**	0.39605
		Without Bpps	0.27696
	Label Encoding Method 2	**With Bpps**	0.38221
		Without Bpps	0.27696
deg_50C	**Label Encoding Method 1**	**With Bpps**	0.32985
		Without Bpps	0.29906
	Label Encoding Method 2	**With Bpps**	0.31735
		Without Bpps	0.29906
Total	**Label Encoding Method 1**	**With Bpps**	0.39575
		Without Bpps	0.33040
	Label Encoding Method 2	**With Bpps**	0.37803
		Without Bpps	0.33040

TABLE IV. PREDICTION ERROR FOR LGBM RESULTED AFTER EVALUATION WITH RMSE PERFORMANCE METRIC.

Classes	Criteria		LGBM
reactivity	**Label Encoding Method 1**	**With Bpps**	0.22215
		Without Bpps	0.13878
	Label Encoding Method 2	**With Bpps**	0.22203
		Without Bpps	0.13878

deg_Mg_pH10	**Label Encoding Method 1**	**With Bpps**	0.28266
		Without Bpps	0.30032
	Label Encoding Method 2	**With Bpps**	0.28272
		Without Bpps	0.30032
deg_pH10	**Label Encoding Method 1**	**With Bpps**	0.26064
		Without Bpps	0.42824
	Label Encoding Method 2	**With Bpps**	0.26083
		Without Bpps	0.42824
deg_Mg_50C	**Label Encoding Method 1**	**With Bpps**	0.23767
		Without Bpps	0.22908
	Label Encoding Method 2	**With Bpps**	0.23759
		Without Bpps	0.22908
deg_50C	**Label Encoding Method 1**	**With Bpps**	0.22019
		Without Bpps	0.24846
	Label Encoding Method 2	**With Bpps**	0.22030
		Without Bpps	0.24846
Total	**Label Encoding Method 1**	**With Bpps**	0.24466
		Without Bpps	0.26897
	Label Encoding Method 2	**With Bpps**	0.24469
		Without Bpps	0.26897

TABLE V. PREDICTION ERROR FOR RF RESULTED AFTER EVALUATION WITH RMSE PERFORMANCE METRIC.

Classes	**Criteria**		**RF**
reactivity	**Label Encoding Method 1**	**With Bpps**	0.23858
		Without Bpps	0.14163
	Label Encoding Method 2	**With Bpps**	0.23871
		Without Bpps	0.14086
deg_Mg_pH10	**Label Encoding Method 1**	**With Bpps**	0.30128
		Without Bpps	0.27372
	Label Encoding Method 2	**With Bpps**	0.30131
		Without Bpps	0.27327
deg_pH10	**Label Encoding Method 1**	**With Bpps**	0.27564
		Without Bpps	0.40627
	Label Encoding Method 2	**With Bpps**	0.27600
		Without Bpps	0.40891
deg_Mg_50C	**Label Encoding Method 1**	**With Bpps**	0.25452
		Without Bpps	0.19598
	Label Encoding Method 2	**With Bpps**	0.25427
		Without Bpps	0.19434
deg_50C	**Label Encoding Method 1**	**With Bpps**	0.23403
		Without Bpps	0.24824
	Label Encoding Method 2	**With Bpps**	0.23394
		Without Bpps	0.24773
Total	**Label Encoding Method 1**	**With Bpps**	0.26081
		Without Bpps	0.25317
	Label Encoding Method 2	**With Bpps**	0.26085
		Without Bpps	0.25302

TABLE VI. MEAN PREDICTION ERROR FOR LR, LGBM, AND RF COMPARED TO MODELS DEVELOPED BY ANKIT SINGHAL (LSTM, GRU, AND GCN) ACROSS ALL 5 PREDICTED CLASSES RESULTED AFTER EVALUATION WITH RMSE PERFORMANCE METRIC.

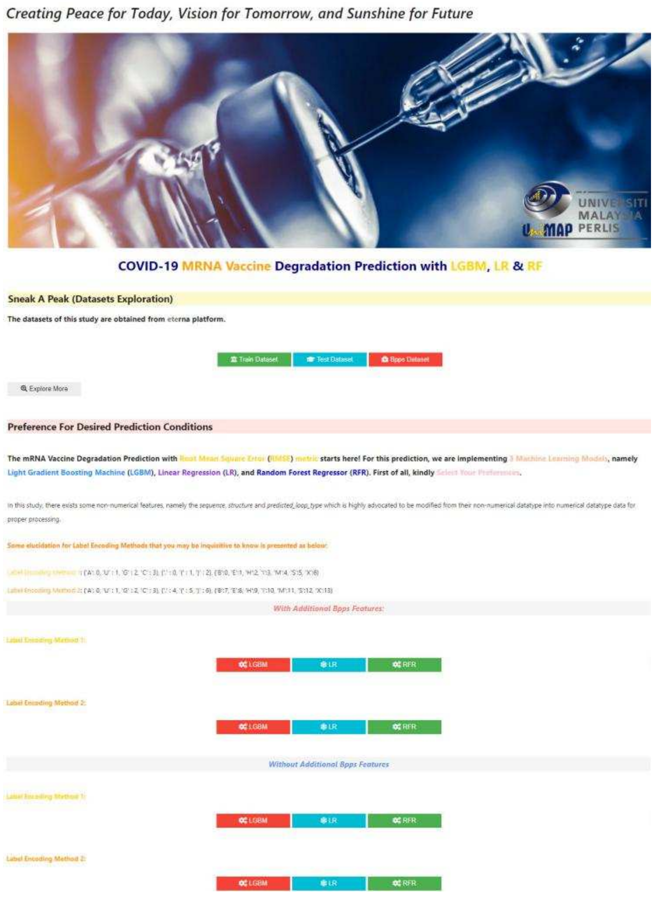
Models	**Criteria**		**Mean RMSE**
LR	**Label Encoding Method 1**	**With Bpps**	0.39575
		Without Bpps	0.33040
	Label Encoding Method 2	**With Bpps**	0.37803
		Without Bpps	0.33040
LGBM	**Label Encoding Method 1**	**With Bpps**	0.24466
		Without Bpps	0.26897
	Label Encoding Method 2	**With Bpps**	0.24469
		Without Bpps	0.26897
RF	**Label Encoding Method 1**	**With Bpps**	0.26081
		Without Bpps	0.25317
	Label Encoding Method 2	**With Bpps**	0.26085
		Without Bpps	0.25302
LSTM	**Hot Encoding**		0.268
GRU	**Hot Encoding**		0.263
GCN	**Hot Encoding**		0.267

From Table 5, with an RMSE value of 0.24466, the LGBM-based model with the interpolate bpps features alongside method 1 label encoding approach outgun the other 11 models. This result also surpassed all the models developed by researcher Ankit Singhal [2]. By the same token, referring to Table 3, 4, and 5, the 6 models that trained with datasets which accrued with bpps features, regardless of its label encoding method, LGBM-based model thrive better result with lower RMSE values, followed by RF-based models and lastly the LR-based model, across all the 5 studied classes. On the other hand, for the remaining 6 models that trained sans additional bpps features, asides from the reactivity class that scored by LGBM with 0.13878 RMSE value for both label encoding methods, RF-based models achieved splendidly for all the other classes despite the exploited label encoding methods, arguably enough to be titled as the titlist model for this mRNA vaccine degradation prediction in default of bpps features compared to LR and LGBM-based models under congruent manner. However, despite the prevailing aspect, it is undoubtedly that across all the 12 machine learning-based models, all in all, the LGBM-based model which trained with datasets incorporated with bpps features and encoded with method 1 label encoding that resulted in a value of 0.24466 after evaluated with RMSE metric predominates the rest for this degradation performance task assigned.

As presented in Table 3, across all the classes, the reactivity, deg_Mg_pH10, deg_pH10, deg_Mg_50C and deg_50C, regardless of whether there is a presence of the auxiliary bpps aggregated-functions-features, reactivity class has a better RMSE result compared to other classes for LR-based algorithm models. This implies that for LR's models, the reactivity class fits the models better and is more accurate to the actual results for degradation predictions, compared to other classes which the prediction error may be slightly larger as imparted by the RMSE values presented. The same goes for LGBM's and RF's models, which demonstrated that among all the 5 classes predicted, prediction results for reactivity will panoply smaller deviation contrary to others, orchestrated in Table 4 and 5. In addition, overall, sussed that other than deg_Mg_pH10 and deg_Mg_50C which is more suitable to be predicted using RF models that implemented label encoding method 2 with the absence of bpps features, the other 3 classes will showcase a better prediction performance with loftier accuracy if predicted using LGBM-based models. Precisely, deg_pH10 and deg_50C class fit more to LGBM-based model where features labelled with label encoding method 1 at one and at the same time under the same nose of bpps features, while reactivity performs least deviation error with LGBM-based models without the escort of bpps aggregated-functions-features.

D. GUI

GUI is a method that allow interfacing and communicating between the user and the system. The GUI of this study is divided into 2 main parts, specifically, the datasets exploration section and the models' prediction evaluation section. The general layout for this system is illustrated in Figure 4.



Creating Peace for Today, Vision for Tomorrow, and Sunshine for Future

COVID-19 MRNA Vaccine Degradation Prediction with LGBM, LR & RF

Sneak A Peak (Datasets Exploration)

The datasets of this study are obtained from eterna platform.

Preference For Desired Prediction Conditions

The mRNA Vaccine Degradation Prediction with Root Mean Square Error (RMSE) metric starts here! For this prediction, we are implementing 3 Machine Learning Models, namely Light Gradient Boosting Machine (LGBM), Linear Regression (LR), and Random Forest Regressor (RFR). First of all, kindly Select Your Preferences.

In this study, there exists some non-numerical features, namely the *sequence*, *structure* and *predicted_loop_type* which is highly advocated to be modified from their non-numerical datatype into numerical datatype data for proper processing.

Some elucidation for Label Encoding Methods that you may be inquisitive to know is presented as below:

Label Encoding Method 1: {'A': 0, 'U' : 1, 'G' : 2, 'C' : 3}, {'.' : 0, '(' : 1, ')' : 2}, {'B':0, 'E':1, 'H':2, 'I':3, 'M':4, 'S':5, 'X':6}

Label Encoding Method 2: {'A': 0, 'U' : 1, 'G' : 2, 'C' : 3}, {'.' : 4, '(' : 5, ')' : 6}, {'B':7, 'E':8, 'H':9, 'I':10, 'M':11, 'S':12, 'X':13}

With Additional Bpps Features:

Label Encoding Method 1:

Label Encoding Method 2:

Without Additional Bpps Features

Label Encoding Method 1:

Label Encoding Method 2:

Fig. 4. General layout of GUI mRNA vaccine degradation prediction tool

The GUI is design with consideration, the prediction performance for all the 12 models evaluated by the RMSE metric could be displayed depending on the user's fascination, on which encoding methods and dataset features catch the user's attention and inquisitiveness the most. With the idea to ease the juxtaposition process among models to ensure the GUI user-friendly specification and at the same time achieve delightful visualization standard while interacting with the user, the GUI prediction tool for this study is outlined to have the ability to display various information on the same interface layout neither jumbling nor interfere with the other sections' and subsections' outputs.

IV. Conclusion

Referring to the result obtained, we may conclude that for this study, LGBM is more congruous than LR and RF for mRNA degradation predictions. LGBM-based model's performance even surpasses the models developed by researcher Ankit Singhal with LSTM, GRU and GCN each with k = 4 ratifying that this highly extolled model is an excellent model to be considered for further modification to be aggrandized for COVID-19 mRNA vaccine degradation prediction study.

On the other hand, ascertained that although RF-based models obscurely underperform the superb mellowed LGBM-based model that trained together with the riddled bpps features, but RF-based models are preponderant in a way that fewer data are required. RF-based model with the implementation of label encoding method 2 and without the support of bpps features that scored splendidly is a good stepping stone for the COVID-19 mRNA vaccine degradation prediction tools modelling as it coincides to the actual situation used to be faced in reality where precise and exuberantly wealth of data is hard to be obtained. In addition, as epitomized in Table 6, under criteria with the absence of annexed bpps features, RF-based models verily outvie the LGBM-based models by 0.01595, which is very much vaster than the gap, nearly double, presented when comparing under the milieu where bpps features are enclosed. Therefore, it is alleged that RF-based model holds great prediction potential and is believed that refinement and amelioration could be done on the RF-based model to reap a deft model for this significative not to mention significant study.

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 [2] 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 researcher Ankit Singhal 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.

Lastly, for GUI aspect, a more interactive along with expedient and competent user interface could be contrived. Test dataset uploading feature may be subsumed and the duration for output display should be shorten by reinforcing the coding. Moreover, the speed of data processing should be upgraded to aggrandize the functionality and quality of the models and GUI prediction tool.

V. References

[1] A. Wadhwa, A. Aljabbari, A. Lokras, C. Foged, and A. Thakur, “Opportunities and Challenges in the Delivery of mRNA-Based Vaccines,” *Pharmaceutics*, vol. 12, no. 2, p. 102, Jan. 2020.

[2] A. Singhal, “Application and Comparison of Deep Learning Methods in the Prediction of RNA Sequence Degradation and Stability,” Nov. 2020.

[3] H. K. Wayment-Steele *et al.*, “Theoretical basis for stabilizing messenger RNA through secondary structure design,” bioRxiv, p. 2020.08.22.262931, 2020.