Exploration of hybrid deep learning algorithms for covid-19 mRNA vaccine degradation prediction system

Soon Hwai Ing [a,1], Azian Azamimi Abdullah [a,2,*], Mohd Yusoff Mashor [a,3], Zeti-Azura Mohamed-Hussein [b,c,4], Zeehaida Mohamed [d,5], Wei Chern Ang [e,f,6]

[a] Faculty of Electronic Engineering Technology, Universiti Malaysia Perlis, Perlis, Malaysia
[b] Centre for Bioinformatics Research, Institute of Systems Biology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
[c] Department of Applied Physics, Faculty of Science & Technology, Universiti Kebangsaan Malaysia, Selangor, Malaysia
[d] School of Medical Sciences, Universiti Sains Malaysia, Kelantan, Malaysia
[e] Clinical Research Centre, Hospital Tuanku Fauziah, Ministry of Health Malaysia, Perlis, Malaysia
[f] Department of Pharmacy, Hospital Tuanku Fauziah, Ministry of Health Malaysia, Perlis, Malaysia
[1] hwaiingsoon@studentmail.unimap.edy.my ; [2] azamimi@unimap.edu.my; [3] yusoff@unimap.edu.my; [4] zeti@ukm.my; [5] zeehaida@usm.my; [6] wei.ang.1990@gmail.com
* corresponding author

ARTICLE INFO

Article history
Selected paper from The 2022 5th International Symposium on Advanced Intelligent Informatics (SAIN'22), Yogyakarta (Virtually), September 14, 2022, http://sain.ijain.org/2022/. Peer-reviewed by SAIN'22 Scientific Committee and Editorial Team of IJAIN journal

Received July 19, 2022
Revised August 24, 2022
Accepted November 30, 2022
Available online November 30, 2022

Keywords
mRNA vaccine
Hybridizing sequence
Degradation
Label encoding
Deep learning algorithms

ABSTRACT

Coronavirus causes a global pandemic that has adversely affected public health, the economy, including every life aspect. To manage the spread, innumerable measurements are gathered. Administering vaccines is considered to be among the precautionary steps under the blueprint. Among all vaccines, the messenger ribonucleic acid (mRNA) vaccines provide notable effectiveness with minimal side effects. However, it is easily degraded and limits its application. Therefore, considering the cruciality of predicting the degradation rate of the mRNA vaccine, this prediction study is proposed. In addition, this study compared the hybridizing sequence of the hybrid model to identify its influence on prediction performance. Five models are created for exploration and prediction on the COVID-19 mRNA vaccine dataset provided by Stanford University and made accessible on the Kaggle community platform employing the two deep learning algorithms, Long Short-Term Memory (LSTM) as well as Gated Recurrent Unit (GRU). The Mean Columnwise Root Mean Square Error (MCRMSE) performance metric was utilized to assess each model's performance. Results demonstrated that both GRU and LSTM are befitting for predicting the degradation rate of COVID-19 mRNA vaccines. Moreover, performance improvement could be achieved by performing the hybridization approach. Among Hybrid_1, Hybrid_2, and Hybrid_3, when trained with Set_1 augmented data, Hybrid_3 with the lowest training error (0.1257) and validation error (0.1324) surpassed the other two models; the same for model training with Set_2 augmented data, scoring 0.0164 and 0.0175 MCRMSE for training error and validation error, respectively. The variance in results obtained by hybrid models from experimenting claimed hybridizing sequence of algorithms in hybrid modeling should be a concerned.

1. Introduction

Treatments for SARS-CoV-2 have caused millions of deaths since the end of 2019, and this pandemic is yet to be invented [1]. Therefore, authoritative vaccines are needed to control the outbreak. Unfortunately, even though mRNA vaccines have shown a promising effect, it has a drawback of rapid degradation.

Wadhwa et al. [2] showed that degradation has significantly reduced the mRNA yields during the in-vitro transcription. Note that the half-life of mRNA vaccines might also be 900 days in cold chain conditions, having a rate of more than 2% degradation every 30 days [2]. Moreover, it is important to note that the half-life of vaccines can be drastically shortened to 5 and 10 days, accordingly, having a temperature digression to about 37°C or with a 2 unit drift with respect to the pKa value [2]. The in-vitro transcription of the vaccines can be conducted at a temperature of 37°C with magnesium ions (Mg2+), which subsequently reduces the half-life to no longer than 2 hours [2]. The outcome, in which the mRNA vaccine is unstable since degradation still happens throughout the transcription process, is similar despite lowering the Mg2+ temperature, or concentration, pH value to reduce the hydrolysis [2].

Besides, Abbasi [3] claimed that this restriction might be circumvented with a second and perhaps booster dose regimen of the vaccine. However, the degradation concern must not be disregarded since a vaccine's effectiveness cannot be replaced or recovered after it has been compromised. Stabilizing mRNA-based vaccines has always been a great challenge. Looking for an optimum solution is like facing an enigma with no end and has caused headaches to vaccine scientists and researchers for decades. Parenthetically, vaccines that become unstable have induced countless losses of lives [4], especially during a pandemic. This study is crucial to address the safety concerns to ensure no adverse impacts on the potency of a vaccine. Vaccine functionality and characteristics are easily affected by a minor degradation [5], and its degradation rate is easily altered by both intrinsic and extrinsic factors.

It is important to research the degradation of the mRNA vaccine. However, few studies were performed on predicting mRNA or vaccine degradation, especially concerning COVID-19 mRNA vaccines. By the end of 2020, research by Singhal on the topic of COVID-19 mRNA vaccine degradation prediction utilizing Graph Convolution Network (GCN), Gated Recurrent Unit (GRU), as well as Long-Short-Term-Memory Cells (LSTM) algorithms assessed with root mean square error (RMSE) revealed that GCN-based model (0.249) is the finest for reactivity prediction. Meanwhile, the GRU-based model with an accuracy of 76% is marked as the premium predictor when considering all the target variables [6].

Imran et al. used a regularized LSTM model to forecast the degradation rate with respect to the mRNA vaccine, and it showed better performance than tree-based algorithms. Different activation functions, including linear, hyperbolic tangent (Tanh) as well as a rectified linear unit (ReLU), were taken into consideration for each layer of the model during model development to converge the Mean Columnwise Root Mean Square Error (MCRMSE) losses [7]. GRU-related models were also proposed and considered [8] [9] [10]. A modified GRU with a multi-head attention mechanism was developed by Wang et al. to train the model by having 3 GCNs to deal with three adjacency matrices, i.e., base-pairing probability (BPP), structure adjacency and distance matrices, respectively. The model performance is measured with MCRMSE, achieving a passable score of 0.3489 [8].

Muneer et al. and Qaid et al. used hybrid models to predict the degradation rate [9] [10], and in tandem with Convolutional Neural Network (CNN), the authors came up with GCN_GRU and GCN_CNN models [9]. Between these models, GCN_GRU pre-trained embedding model showed the best performance with a score of 0.938 for the Area Under the Curve (AUC) performance metric, which indicated its suitability in the base-wise reactivity prediction studies compared to CNN. On the other hand, the three models proposed by Qaid et al. include LSTM, GRU, and hybrid LSTM_GRU. Different from other research, the authors suggested two different encoding methods, i.e., the base (0 – 13) encoding method and the codon (1 – 434) method of encoding. Results presented that LSTM trained with the codon encoding method is the best model among all the proposed models [10]. However, the authors suggested that the base encoding method is much more preferable compared to the codon encoding method since it has a lesser tendency to overfitting.

Other than deep learning algorithms, Ing et al. proposed three machine learning algorithms (Random Forest, Light Gradient Boosting Machine, as well as Linear Regression) with respect to this prediction study [11] [12], showing that both theories of machine learning and deep learning are comfortable with this study. Referring to all the past studies reviewed, it is found that GRU and LSTM are the two widely

used algorithms in this field of research. Besides, it deduced that hybridizing algorithms to form a hybrid model for prediction is conducive to reducing the error. However, it is noticed that researchers failed to demonstrate the results of hybrid models concerning a different sequence of hybridization. Hence, this paper presents the prediction results of hybrid models considering the hybridizing sequence with GRU and LSTM algorithms, utilizing the concept proposed by Qaid et al. in [10].

2. Method

Besides developing reliable models that have the ability to forecast the rate of the COVID-19 mRNA vaccine degradation, this paper also focuses on discovering the relationship between predicted results with the hybridizing sequence of algorithms in hybrid modeling. To ensure comparability, this research utilized the same datasets and performed the same concept and theory as executed by Qaid et al., except excluding the codon encoding method. Since researchers' main objective is to develop degradation prediction models with absolute accuracy with low error rates, only the training and bpps datasets are extracted from the Kaggle community and Eterna platforms. This was bolstered by Stanford University [13] to perform a supervised-based study instead of a semi-supervised. Several features in the forms of aggregate functions (Exponential Weighted Average, Maximum, Normalize, Average Position Value, and Summation) were engineered from the BPPs dataset to represent the numerical features. The pre-processing step handles eliminating noises and data organization for training and evaluating purposes. Completion of the pre-processing data stage will generate well-encoded, clean data that is ready to be fed into models, followed by the training of 5 models with the trained dataset for model development. Model performance was evaluated on the validation set with MCRMSE. Fig. 1 illustrates the method workflow.

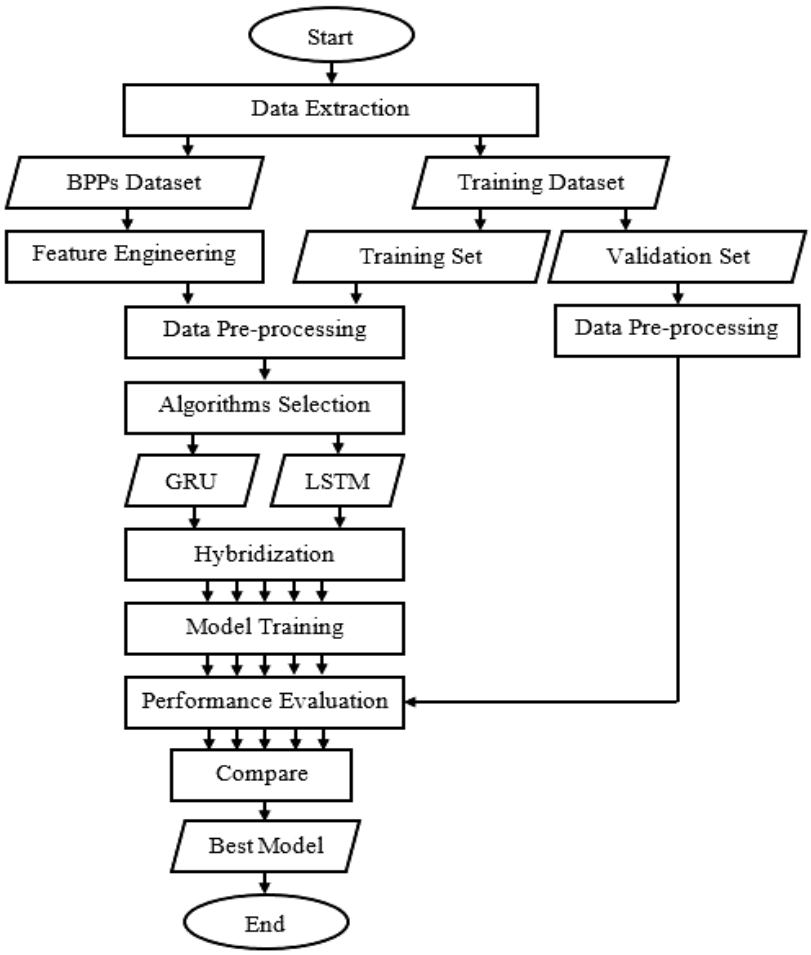

Fig. 1. The Flow Chart of the Methodology

2.1. Dataset

The empirical data obtained from Kaggle sourced by Stanford University [14] for this research gave rise to a regression study, allowing the prediction of the degradation rate of the mRNA vaccine to be studied. Together with the augmented dataset and at the same time having the training dataset contributing 2400 samples to the total amounts, two sets of data, one with 4800 samples and another with 21600 samples, are gained.

2.1.1. Training Dataset

The training dataset extracted comprises 2400 samples with 19 fields, inclusive of the '*sequence*', '*structure*', '*predicted_loop_type*', '*SN_filter*', '*signal_to_noise*', '*seq_length*', '*seq_scored*', '*reactivity_error*', '*deg_error_pH10*', '*deg_error_Mg_pH10*', '*deg_error_50C*', '*deg_error_Mg_50C*', '*reactivity*', '*deg_50C*', '*deg_pH10*', '*deg_Mg_pH10*', '*deg_Mg_50C*', comprising the '*id*' and '*index*'. Among these 19 fields, '*sequence*', '*structure*', as well as '*predicted_loop_type*' are the three input fields while '*reactivity*', '*deg_50C*', '*deg_pH10*', '*deg_Mg_pH10*', and '*deg_Mg_50C*' will be the output fields. A 90:10 percentage split is performed on the training dataset resulting in two sets of data, the training set that holds 90% of samples obtained from the initial training dataset as well as the validation set that holds the remaining 10%. The training set will be used for models' training, while the validation set will be merited during performance evaluation. List of features show in Table 1.

Table 1. List of Features

Features	Description
index	Numerical order list with respect to each sample
id	An identifier with respect to each sample
sequence	Combination of A, U, G, and C bases for each sample, depicting the RNA sequence.
structure	Alignment of '(', ')' and '.' characters, illustrating the pairing state of the RNA. Here, '(' and ')' indicate paired base, while '.' indicate unpaired interaction. Length correlates to the *'sequence'* feature.
predicted_loop_type	Describe the structure of each character in '*sequence*' that delineates RNA structures having code name called 'bpRNA'. B: Bulge; H: Hairpin loop; E: dangling End; I: Internal loop; S: paired "Stem"; X: eXternal loop; M: Multiloop
signal_to_noise	Determine the quality of the sample. Higher SNR, better quality.
SN_filter	Denoted with 1 if the sample fullfils both 2 filter conditions; else 0. 2 conditions, taking into consideration that only the first 68 bases of RNA samples sequence in the Train dataset: (1) Minimum value > – 0.5 across all 5 outputs. (2) Mean SNR > 1.0 across all 5 outputs.
seq_length	Depict the length of the '*sequence*' of RNA samples, 107.
seq_scored	68, depicting the number of positions utilized in scoring with predicted values, is analogous to the length of all 5 classes together with their 'error'.
reactivity_error	A series of calculated errors arising from experimenting with respect to '*reactivity*'
deg_error_Mg_pH10	A series of calculated errors arising from experimenting with respect to '*deg_Mg_pH10*.
deg_error_pH10	A series of calculated errors arising from experimenting with respect to '*deg_pH10*'.
deg_error_Mg_50C	A series of calculated errors arising from experimenting with respect to '*deg_Mg_50C*'.
deg_error_50C	A series of calculated errors arising from experimenting with respect to '*deg_50C*'.
reactivity	A series of numbers conveying the probability of the base being paired.
deg_Mg_pH10	A series of numbers conveying the fragility of the linkage in each base under pH10, with the occurrence of Mg.
deg_pH10	A series of numbers conveying the fragility of the linkage in each base under pH10 having no occurrence of Mg.
deg_Mg_50C	A series of numbers conveying the fragility of the linkage in each base under 50°C, with the occurrence of Mg.
deg_50C	A series of numbers conveying the fragility of the linkage in each base under 50°C, having no occurrence of Mg.

2.1.2. Augmented Dataset

Deep learning algorithms have an innate defect of tending to overfit the data [15] [16] [17]. To avoid the overfitting issue, analysts suggest increasing the number of training samples to ensure diversion. Still, data collection requires a lot of procedure and resources and is undeniably time-consuming. Therefore, augmenting existing data by modifying the samples will usually be the preference for most practitioners in circumventing overfitting [18] [19]. For this research, we utilized two different sets of augmented data. The first set of augmented data (Set_1) is the attached augmented data generated with the ARNIE package offered by Qaid et al. [10] to ensure utter comparability is attained. In contrast, the second set of augmented data (Set_2) is a public dataset retrieved from [20].

2.1.3. BPPs Dataset

Kaggle platform presented a set of data that comprised 6034 BPPs symmetric square matrix NumPy file in forming the BPPs dataset. Summing the 2400 samples from the training dataset and 3634 samples from the testing dataset resulted in the amount of 6034 BPPs files. However, this research focused on supervised learning, utilizing only the training samples. This BPPs dataset will be engineered to generate useful aggregate function features, also known as the numerical features, by Qaid et al.

2.2. Data Pre-processing

Fallacious, abominable, or nugatory data will alter the prediction accuracy and quality [21] [22] [23] [24]. Therefore, to eschew undesired complications, cleaning and simplifying noisy, crude data to ease data handling and minimizing the reduction of data quality by conducting procedures of data pre-processing is a crucial step.

2.2.1. Data Cleaning

Practicing the first phase of exploratory data analysis with the '.isnull' command, discovered no missing value in the dataset extracted; however, the '*signal_to_noise*' field uncovered that the dataset subsumed noisy samples. Therefore, the dataset is filtered with stipulated *SN_filter* criteria as proclaimed in Table 1 to ensure solely refined samples are preserved. After filtering, a total of 304 noisy samples are removed from the training dataset

2.2.2. Label Encoding

Data could come in multiple data types, i.e., categorical, ordinal and numerical, but recommended to be modified into numerical since some algorithms that could not manage non-numerical data exist [25]. Label encoding is suggested to encode the three non-numerical inputs: '*predicted_loop_type*,' '*structure*,' as well as '*sequence*.' Here, the characters are base encoded as depicted in Table 2.

Table 2. Label Encoding

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

2.3. Feature Engineering

The quality of inputs, also known as features, will determine the aptitude of a model. Processing time and storage space can be greatly saved with the presence of first-string quality inputs. To process the raw BPPs matrix dataset into a more apposite form of inputs, the dataset is feature engineered into a quinary of aggregator-function inputs, that is, the numerical features introduced in [10], i.e., Exponential Weighted Average, Maximum, Normalize, Position Average Value and Summation. Ing et al. have proven to carry data visualization techniques in determining the suitability and safety of

engineered numerical features [11] [12]. However, since this study engaged only the training dataset in which all the samples hold an equal number of bases for each sequence, referred to as '*seq_length*,' this research harnessed all the numerical features as inputs.

2.4. Deep Learning

GRU and LSTM are the two deep learning algorithms implemented for this degradation rate prediction study. The five models developed with these two algorithms are evaluated with the MCRMSE performance metric.

2.4.1. Gated Recurrent Unit (GRU)

GRU may be presented as a spinoff of LSTM [26], a type of RNN. Although GRU is lucid and more compact than LSTM, not only the competency in mastering context is not omitted, but on the contrary, reducing the training time [27] [28]. Alluded to research conducted by [6] as well as [8] [9] [10] on predicting COVID-19 mRNA vaccine degradation rate, it is deduced that GRU is indeed an applicable algorithm for this bioinformatics-related artificial intelligence-based research.

2.4.2. Long Short-Term Memory (LSTM)

Compared to GRU, which has only two gates (update gate as well as reset gate) in modulating information flow, LSTM has higher gates (output gate, forget gate, as well as input gate) for information winnowing [29] [30] [31], leading LSTM to have a higher complexity but better accuracy than GRU. If accuracy was of priority and a large dataset was practiced, LSTM used to be cherry-picked by researchers more than GRU. The off-the-rack results presented in [10] by Qaid et al. have had this argument testified.

2.5. MCRMSE Performance Metric

The performance and effectiveness of the proposed model will then be evaluated with a performance metric. This study is a regression related-study that aims to forecast the mRNA vaccine degradation rate with respect to COVID-19. Therefore, regression error is analyzed to study the models' prediction performance, and MCRMSE, which stands for Mean Column-wise Root Mean Squared Error, is proposed. The square root of the mean of the squared variations between the predictions and the ground truth is factored by the regression performance metric known as RMSE to determine the average magnitude of errors [32] [33]. The RMSE metric formula is provided in (1), in which n denotes the number of occurrences.

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

Meanwhile, MCRMSE can be deduced as an average across all RMSE values for each predicted target to obtain an individual number evaluation metric from multiple outputs. The formula for MCRMSE is presented in (2), where N_t will be inputted with the number of targets for prediction scoring.

$$MCRMSE = \frac{1}{N_t}\sum_{j=1}^{N_t}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(actual_{ij} - predicted_{ij})^2}. \tag{2}$$

Equations (1) and (2) attested that both RMSE and MCRMSE are negative-oriented scoring techniques. Graced with the presence of a square in equations, the error ranged from zero to positive infinity

3. Results and Discussion

Other than determining and developing models concerning COVID-19's mRNA vaccines' degradation rate prediction, this paper concentrates on discovering the sequence of hybridizing effects on the prediction results. MCRMSE is engaged as the performance metric for models' prediction performance evaluation across the five outputs.

3.1. Hybridization

This is in contrast to several ensemble approaches with algorithms to serve independently to produce several outcomes followed by polling systems like max voting, weighing, averaging and determining a single final result. On the other hand, hybridization has algorithms that serve dependently to produce a single result with no polling system involved [34].

This research utilized the approach suggested in [10] GRU and LSTM in model development. Taking the three models proposed by Qaid et al. with the additional two hybrid models suggested in this research for hybridization sequence exploration, a total of five models are engaged for this study. All the models constituted three bidirectional layers, with each direction having 256 hidden layers. The first and second models, i.e., the GRU and LSTM models, have all three layers congregated with GRU and LSTM, respectively. The remaining three models were made up of two GRU and one LSTM. The hybrid models are given their appellation name dependent on the layer where the LSTM is occupied. For illustrative purposes, if the hybrid model has had LSTM occupy the first layer and the remaining two layers by GRU, it is named Hybrid_1. The hybridizing sequence of each model is detailed in Table 3.

Table 3. The sequence of Hybridization of Algorithms of Each Developed Model.

Model	Sequence		
	Bidirectional Layer 1	*Bidirectional Layer 2*	*Bidirectional Layer 3*
GRU	GRU	GRU	GRU
LSTM	LSTM	LSTM	LSTM
Hybrid_1	LSTM	GRU	GRU
Hybrid_2	GRU	LSTM	GRU
Hybrid_3	GRU	GRU	LSTM

.

3.2. Prediction Performance

The overall filtered dataset is classified into a validation set as well as a training set with a 90:10 percentage split to ensure this research achieves utter comparability with Qaid et al.'s research. When the augmented data is not included, the split produces training and validation data of 1886 and 210 samples, respectively. However, when Set_1 augmented data is considered, 3772 samples of training data and 420 samples of validation data are split from the 4192 samples of the training dataset that have had 608 noisy samples removed. In addition, when Set_2 is involved, removing 3076 noisy samples from 21600 samples, the remaining 18524 samples in the training data are divided into validation data with 1853 samples as well as training data with 16671 samples for model development. That aside, several parameters and hyperparameters are initialized as tabulated in Table 4 to configure the models.

Table 4. Parameters And Hyperparameters Initialization For Models Configuration.

Parameters/ Hyperparameters	Value
Activation Function	Linear
Dropout	0.5 or 0
Embedding Layer	100
Hidden Units	256
Kernel Initializer	Orthogonal
Layers	Bidirectional
Optimizer	Adam
Return Sequence	True
Test Size	0.1

Table 5 shows the deep learning models' prediction performance evaluated with MCRMSE.

Table 5. Prediction Error For Developed Model Obtained After Evaluation With MCRMSE Performance Metric.

Dropout	Augmented Data	BPPs	Data	Model				
				GRU	*LSTM*	*Hybrid_1*	*Hybrid_2*	*Hybrid_3*
0.5	Without	Without	Training	0.1582	0.1378	0.1598	0.1551	0.1530
			Validation	0.2124	0.2144	0.2130	0.2137	0.2138
		With	Training	0.1551	0.1498	0.1520	0.1493	0.1577
			Validation	0.2109	0.2138	0.2107	0.2101	0.2127
	Set _1	Without	Training	0.1375	0.1178	0.1349	0.1295	0.1269
			Validation	0.1426	0.1289	0.1392	0.1358	0.1352
		With	Training	0.1360	0.1180	0.1331	0.1299	0.1257
			Validation	0.1400	0.1278	0.1376	0.1345	0.1324
	Set_2	Without	Training	0.0961	0.0755	0.0920	0.0878	0.0901
			Validation	0.0800	0.0594	0.0766	0.0714	0.0744
		With	Training	0.0976	0.0760	0.0845	0.0867	0.0914
			Validation	0.1156	0.0598	0.0786	0.0754	0.0957
0	Set_2	Without	Training	0.0258	0.0156	0.0225	0.0204	0.0200
			Validation	0.0291	0.0174	0.0249	0.0227	0.0235
		With	Training	0.0214	**0.0143**	0.0200	0.0178	**0.0164**
			Validation	0.0220	**0.0151**	0.0206	0.0180	**0.0175**

Referring to Table 5, it is observed that although the overall results obtained are slightly better than the results presented by [10], the ±0.005 difference is too paltry to be considered when compared with the gained loss errors. Nevertheless, this study addresses the effects of hybridizing sequence in the model on the prediction performance with the mRNA vaccines rate of degradation dataset. Meanwhile, probing the contribution of the numerical BPPs inputs to the prediction.

From Table 5, regardless of the presence or absence of the BPPs numerical inputs when the dropout value is set to 0.5 and involves Set_1 augmented data, the LSTM model scored better than the other four models. However, when the Set_1 augmented data is not committed in the experiment, although haunted with overfitting issues, hybrid models have shown lower error rates than the LSTM model. Setting the dropout value to 0.5, even though the MCRMSE loss of the LSTM model (0.1378) on training data is much lower than the other four models when both Set_1 augmented data and numerical BPPs inputs are absent, its validation error loss is the highest. In short, a deduction on the LSTM model can outshine the other four models when interacting with conversant samples but not with unacquainted samples that can be drawn from these results under the criteria. These results indicate that even if the overfitting issue is lifted, the LSTM model may not be qualified as the wistful model to be considered.

Worth noting that when no augmented data is involved, although overfit, the hybrid models show lower loss errors than the GRU model (constituting three bidirectional GRU layers) and the LSTM model (which comprises three bidirectional LSTM layers). This result presents that hybridization is indeed practicable for better model performance at the same time, showing the claim that LSTM can achieve better performance than GRU is only applicable when big data is involved.

This research involved two sets of augmented data to study the hybridization sequence of models for predicting the degradation rate of the mRNA vaccine. Besides Set_1, the prediction errors of generated models trained with Set_2 augmented data are also available in Table 5. It is discovered that when the dropout value is set to 0.5, all models are wiped out by an underfitting issue when trained with the Set_2 augmented data without the presence of BPPs numerical inputs. The results have presented that when models are trained with Set_2 augmented data, involving numerical inputs is no better than excluding them. Observation from the prediction errors tabulated in Table 5 discovered that, besides the GRU model and Hybrid_3 model, all the remaining three models are being whipped.

Valuing dropout with 0.5, when Set_1 augmented data is engaged, no overfitting nor underfitting issue arises, but when engaging model training with Set_2, virtually all models face an underfitting issue. Therefore, to allow the proceeding of the research, the dropout value is tuned to zero and experimented

with Set_2 augmented data on all the five generated models. Dropout is a class of stochastic approaches introduced originally by Hinton et al. [35] to be employed in practice, such as regularisation, model compression, handling overfitting, and more [36] [37] [38]. Tuning the dropout value has outlined its ability to handle underfitting besides solving overfitting.

Dropout is a process involving neurons of a neural network [39] [40], while neurons can be described as some weight-linked processors [41] [42]. Weights and activation functions are the two main components in neurons besides inputs and outputs [43] [44], but the number of neurons is arbitrary. There are no specific rules for prior determination of the number of neurons occurring in each layer with respect to a model . The number of neurons will determine the degree of complexity of a model [45]. Although overfitting can be solved by dropping some neurons [46], dropping out too many neurons will induce underfitting, like those results when trained with Set_2 augmented data with dropout value 0.5 shown in Table 5.

After assigning zero to the dropout value, prediction errors show that all the models manage to have better performance with the presence of BPPs inputs when Set_2 augmented data is involved. Again, the LSTM model surpasses the other models by scoring and achieving the lowest error rate. Meanwhile, among Hybrid_1, Hybrid_2, as well as Hybrid_3, it seems that Hybrid_3 possesses the lowest errors and manages to rank second, in tow to the LSTM model. With the difference in prediction errors scored by these three hybrid models, the message that delivers the importance of hybridizing sequence is once again stressed.

Moreover, as observed from the loss errors tabulated in Table 5, taking the numerical inputs engineered from the BPPs dataset alone does not solve overfitting or underfitting problems. In virtue of augmented data, it is observed that the numerical features have improved the performance trivially by reducing the errors by at most 0.002 with Set_1 augmented data and 0.004 with Set_2 augmented data. However, focusing on Set_1 augmented data, surprisingly, LSTM fits better with the dataset without the numerical inputs. Although the effect is weeny, the numerical inputs bring no good impact to the LSTM model. Even with Set_2 augmented data, although the presence of BPPs manages to help in reducing the error, the improvement is merely just ±0.002 compared to without it. With the results, reconsideration on implementation of numerical features that show low competency (±0.002 or ±0.004) that is too pittance to be discerned compared to the losses error is required when taking computational time and complicity into consideration.

Among the hybrid models, when augmented data is involved, regardless if the augmented data is Set_1 or Set_2, results show that the Hybrid_3 model performed better than Hybrid_2, followed by Hybrid_1. When there is a presence of numerical inputs but an absence of augmented data, Hybrid_2 scored better than Hybrid_1 and Hybrid_3 when BPPs numerical inputs are considered; but, Hybrid_3 has a lower prediction error when both augmented data and BPPs numerical inputs are absent. These results have proven that both the training factors and the sequence of hybridizing algorithms in model formation influences prediction performances.

4. Conclusion

Referring to the results obtained, it may be established that both GRU and LSTM are applicable for this mRNA vaccine's degradation rate prediction research. Notice that when the data augmentation process is not practiced, the overfitting issue is more severe in the LSTM model than in all the other developed models. But, when the sample size is doubled or more, the LSTM model outdid the other models, proving that LSTM is more suited for big data prediction. Over and above that, theorized that achieving a good result can only be granted if the complexity of the model is in jibed with the dataset.

Better pattern recognition and easier model fitting to the dataset can be achieved with fine features and inputs. Still, it is essential to be prudent with the implementation of additional engineered features. For example, suppose the features show no promising merits to the model in prediction performance. In that case, it is recommended to exclude the features as inputs for model training as they will magnify

the intricacy of models and lengthen the computation time, which is pyrrhic. The results in Table 5 have validated the argument that hybridization is a good approach for performance improvement. Furthermore, the difference in results presented by Hybrid_1, Hybrid_2, and Hybrid_3 attested to the claim that the prediction performance of a hybrid model is not solely dependent on the factors in the training stage but also on the sequence of algorithms being hybridized for model development. Therefore, experimentation, along with trial and error, is required to examine the sequencing effect of the algorithms with respect to the performance involving hybridization. As concluded, the results obtained construed that doubling the amount of the original samples resolved the overfitting predicament, highlighting that increasing the amount could further improve the prediction performance by reducing the loss errors. However, further increasing the sample size could burden a model, and underfitting will be induced if the model cannot afford the complexity of the data. Therefore, multiplying the amount of sample is hereupon recommended for future research. Still, at the same time, it should never overlook the compatibility between model and data to avoid both underfitting and overfitting issues. Moreover, this study only compared the hybrid models suggested in [10]. However, LSTM surpasses GRU in accuracy with its complexity, justified by the results tabulated. Hence, we suggest replacing one of the bidirectional GRU layers with a bidirectional LSTM layer along with hyperparameter tuning to improve the hybrid models proposed by [10]. Furthermore, for future work, it is proposed to hybridize other machine learning models with deep learning models to lessen the complexity of a hybrid model but ascent the prediction performance.

Acknowledgment

This study obtained financial aid from the Ministry of Higher Education Malaysia (FRGS/1/2021/TKO/UNIMAP/02/65).

Declarations

Author contribution. All authors equally provided a significant contribution to this research. The final manuscript was reviewed and approved by all authors.
Funding statement. This research is funded by the Ministry of Higher Education Malaysia (FRGS/1/2021/TKO/UNIMAP/02/65).
Conflict of interest. The authors affirm that they have no known financial or interpersonal conflicts that would have an impact on the research presented in this study.
Additional information. No additional information is relevant for this paper.

References

[1] N. Zhu *et al.*, "A Novel Coronavirus from Patients with Pneumonia in China, 2019," *N. Engl. J. Med.*, vol. 382, no. 8, pp. 727–733, Feb. 2020, doi:10.1056/NEJMoa2001017.

[2] 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, doi: 10.3390/pharmaceutics12020102.

[3] J. Abbasi, "COVID-19 and mRNA Vaccines—First Large Test for a New Approach," *JAMA*, vol. 324, no. 12, p. 1125, Sep. 2020, doi: 10.1001/jama.2020.16866.

[4] N. Dumpa *et al.*, "Stability of Vaccines," *AAPS PharmSciTech*, vol. 20, no. 2, pp. 1–11, Feb. 2019, doi: 10.1208/s12249-018-1254-2.

[5] D. J. A. Crommelin, T. J. Anchordoquy, D. B. Volkin, W. Jiskoot, and E. Mastrobattista, "Addressing the Cold Reality of mRNA Vaccine Stability," *J. Pharm. Sci.*, vol. 110, no. 3, pp. 997–1001, Mar. 2021, doi: 10.1016/j.xphs.2020.12.006.

[6] A. Singhal, "Predicting Hydroxyl Mediated Nucleophilic Degradation and Molecular Stability of RNA Sequences through the Application of Deep Learning Methods," Nov. 2020, Accessed: Jan. 10, 2020. [Online]. Available: http://arxiv.org/abs/2011.05136.

[7] S. Asif Imran, M. Tariqul Islam, C. Shahnaz, M. Tafhimul Islam, O. Tawhid Imam, and M. Haque, "COVID-19 mRNA Vaccine Degradation Prediction using Regularized LSTM Model," in *2020 IEEE*

International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE), Dec. 2020, pp. 328–331, doi: 10.1109/WIECON-ECE52138.2020.9398044.

[8] Y. Wang, "Predicting the Degradation of COVID-19 mRNA Vaccine with Graph Convolutional Networks," in *2021 6th International Conference on Machine Learning Technologies*, Apr. 2021, pp. 111–116, doi: 10.1145/3468891.3468907.

[9] A. Muneer, S. M. Fati, N. Arifin Akbar, D. Agustriawan, and S. Tri Wahyudi, "iVaccine-Deep: Prediction of COVID-19 mRNA vaccine degradation using deep learning," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 9, pp. 7419–7432, Oct. 2022, doi: 10.1016/j.jksuci.2021.10.001.

[10] T. S. Qaid, H. Mazaar, M. S. Alqahtani, A. A. Raweh, and W. Alakwaa, "Deep sequence modelling for predicting COVID-19 mRNA vaccine degradation," *PeerJ Comput. Sci.*, vol. 7, pp. 1–21, Jun. 2021, doi: 10.7717/peerj-cs.597.

[11] S. H. Ing, A. A. Abdullah, N. H. Harun, and S. Kanaya, "COVID-19 mRNA Vaccine Degradation Prediction Using LR and LGBM Algorithms," *J. Phys. Conf. Ser.*, vol. 1997, no. 1, p. 012005, Aug. 2021, doi: 10.1088/1742-6596/1997/1/012005.

[12] S. H. Ing, A. A. Abdullah, and S. Kanaya, "Development of COVID-19 mRNA Vaccine Degradation Prediction System," in *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Sep. 2021, pp. 449–454, doi: 10.1109/3ICT53449.2021.9582052.

[13] H. K. Wayment-Steele *et al.*, "Theoretical basis for stabilizing messenger RNA through secondary structure design," *Nucleic Acids Res.*, vol. 49, no. 18, pp. 10604–10617, Oct. 2021, doi: 10.1093/nar/gkab764.

[14] "OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction | Kaggle." Accessed Jan. 10, 2020, Available : https://www.kaggle.com/competitions/stanford-covid-vaccine/data.

[15] M. Marouf *et al.*, "Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks," *Nat. Commun.*, vol. 11, no. 1, p. 166, Jan. 2020, doi: 10.1038/s41467-019-14018-z.

[16] T. Li, R. Zuo, X. Zhao, and K. Zhao, "Mapping prospectivity for regolith-hosted REE deposits via convolutional neural network with generative adversarial network augmented data," *Ore Geol. Rev.*, vol. 142, p. 104693, Mar. 2022, doi: 10.1016/j.oregeorev.2022.104693.

[17] C. F. G. Dos Santos and J. P. Papa, "Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 1–25, Jan. 2022, doi: 10.1145/3510413.

[18] V. Lalitha and B. Latha, "A review on remote sensing imagery augmentation using deep learning," *Mater. Today Proc.*, vol. 62, pp. 4772–4778, Jan. 2022, doi: 10.1016/j.matpr.2022.03.341.

[19] Y. Wang, S. Luo, and H. Wu, "Defect detection of solar cell based on data augmentation," *J. Phys. Conf. Ser.*, vol. 1952, no. 2, p. 022010, Jun. 2021, doi: 10.1088/1742-6596/1952/2/022010.

[20] "GRU+LSTM with 48k augmentation | Kaggle." https://www.kaggle.com/code/mathurinache/gru-lstm-with-48k-augmentation/data (accessed Jan. 10, 2023).

[21] T. Bayrak and H. Ogul, "Data Integration for gene expression prediction," in *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Sep. 2018, pp. 1–6, doi: 10.1109/IDAP.2018.8620915.

[22] Asniar, N. U. Maulidevi, and K. Surendro, "SMOTE-LOF for noise identification in imbalanced data classification," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 6, pp. 3413–3423, Jun. 2022, doi: 10.1016/J.JKSUCI.2021.01.014.

[23] H. Sun, J. Sun, K. Zhao, L. Wang, and K. Wang, "Data-Driven ICA-Bi-LSTM-Combined Lithium Battery SOH Estimation," *Math. Probl. Eng.*, vol. 2022, pp. 1–8, Mar. 2022, doi: 10.1155/2022/9645892.

[24] X. B. Jin, W. T. Gong, J. L. Kong, Y. T. Bai, and T. L. Su, "A Variational Bayesian Deep Network with Data Self-Screening Layer for Massive Time-Series Data Forecasting," *Entropy 2022, Vol. 24, Page 335*, vol. 24, no. 3, p. 335, Feb. 2022, doi: 10.3390/E24030335.

[25] P. Cheng, J. Wang, X. Zeng, P. Bruniaux, and X. Tao, "Motion comfort analysis of tight-fitting sportswear from multi-dimensions using intelligence systems," *Text. Res. J.*, vol. 92, no. 11–12, pp. 1843–1866, Jun.

2022, doi: 10.1177/00405175211070611.

[26] Z. Qu, L. Su, X. Wang, S. Zheng, X. Song, and X. Song, "A Unsupervised Learning Method of Anomaly Detection Using GRU," in *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Jan. 2018, pp. 685–688, doi: 10.1109/BigComp.2018.00126.

[27] R. Nassif and M. W. Fahkr, "Supervised Topic Modeling Using Word Embedding with Machine Learning Techniques," in *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*, Feb. 2020, pp. 1–6, doi: 10.1109/AECT47998.2020.9194177.

[28] H. Yang and S. Liu, "Water quality prediction in sea cucumber farming based on a GRU neural network optimized by an improved whale optimization algorithm," *PeerJ Comput. Sci.*, vol. 8, p. e1000, May 2022, doi: 10.7717/peerj-cs.1000.

[29] C. Hu, S. Martin, and R. Dingreville, "Accelerating phase-field predictions via recurrent neural networks learning the microstructure evolution in latent space," *Comput. Methods Appl. Mech. Eng.*, vol. 397, p. 115128, Jul. 2022, doi: 10.1016/j.cma.2022.115128.

[30] H. Maru, T. Chandana, and D. Naik, "Comparitive study of GRU and LSTM cells based Video Captioning Models," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Jul. 2021, pp. 1–5, doi: 10.1109/ICCCNT51525.2021.9579565.

[31] N. Zafar, I. U. Haq, J. U. R. Chughtai, and O. Shafiq, "Applying Hybrid Lstm-Gru Model Based on Heterogeneous Data Sources for Traffic Speed Prediction in Urban Areas," *Sensors 2022, Vol. 22, Page 3348*, vol. 22, no. 9, p. 3348, Apr. 2022, doi: 10.3390/S22093348.

[32] C. Padubidri, A. Kamilaris, S. Karatsiolis, and J. Kamminga, "Counting sea lions and elephants from aerial photography using deep learning with density maps," *Anim. Biotelemetry*, vol. 9, no. 1, pp. 1–10, Dec. 2021, doi: 10.1186/s40317-021-00247-x.

[33] V. Venugopal, J. Joseph, M. V. Das, and M. K. Nath, "DTP-Net: A convolutional neural network model to predict threshold for localizing the lesions on dermatological macro-images," *Comput. Biol. Med.*, vol. 148, p. 105852, Sep. 2022, doi: 10.1016/j.compbiomed.2022.105852.

[34] C. F. Tsai and M. L. Chen, "Credit rating by hybrid machine learning techniques," *Appl. Soft Comput.*, vol. 10, no. 2, pp. 374–380, Mar. 2010, doi: 10.1016/J.ASOC.2009.08.003.

[35] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," Jul. 2012, doi: 10.48550/arXiv.1207.0580.

[36] A. Labach, H. Salehinejad, and S. Valaee, "Survey of Dropout Methods for Deep Neural Networks," Apr. 2019, doi: 10.48550/arXiv.1904.13310.

[37] H. Alsayadi, A. Abdelhamid, I. Hegazy, and Z. Taha, "Data Augmentation for Arabic Speech Recognition Based on End-to-End Deep Learning," *Int. J. Intell. Comput. Inf. Sci.*, vol. 21, no. 2, pp. 50–64, Jul. 2021, doi: 10.21608/ijicis.2021.73581.1086.

[38] A. Zunino, S. A. Bargal, P. Morerio, J. Zhang, S. Sclaroff, and V. Murino, "Excitation Dropout: Encouraging Plasticity in Deep Neural Networks," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1139–1152, Apr. 2021, doi: 10.1007/s11263-020-01422-y.

[39] S. Z. K. Tan, R. Du, J. A. U. Perucho, S. S. Chopra, V. Vardhanabhuti, and L. W. Lim, "Dropout in Neural Networks Simulates the Paradoxical Effects of Deep Brain Stimulation on Memory," *Front. Aging Neurosci.*, vol. 12, p. 273, Sep. 2020, doi: 10.3389/fnagi.2020.00273.

[40] Y. Chen and Z. Yi, "Adaptive sparse dropout: Learning the certainty and uncertainty in deep neural networks," *Neurocomputing*, vol. 450, pp. 354–361, Aug. 2021, doi: 10.1016/j.neucom.2021.04.047.

[41] S. Chen *et al.*, "Rainfall Forecasting in Sub-Sahara Africa-Ghana using LSTM Deep Learning Approach," *Int. J. Eng. Res. Technol.*, vol. 10, no. 3, pp. 464–470, Apr. 2021, [Online]. Available: https://www.ijert.org/rainfall-forecasting-in-sub-sahara-africa-ghana-using-lstm-deep-learning-approach .

[42] A. Malekian and N. Chitsaz, "Concepts, procedures, and applications of artificial neural network models in streamflow forecasting," in *Advances in Streamflow Forecasting*, Elsevier, 2021, pp. 115–147, doi: 10.1016/B978-0-12-820673-7.00003-2.

[43] Y. O. Ouma, C. O. Okuku, and E. N. Njau, "Use of Artificial Neural Networks and Multiple Linear Regression Model for the Prediction of Dissolved Oxygen in Rivers: Case Study of Hydrographic Basin of River Nyando, Kenya," *Complexity*, vol. 2020, pp. 1–23, May 2020, doi: 10.1155/2020/9570789.

[44] S. Tiryaki and A. Aydın, "An artificial neural network model for predicting compression strength of heat treated woods and comparison with a multiple linear regression model," *Constr. Build. Mater.*, vol. 62, pp. 102–108, Jul. 2014, doi: 10.1016/j.conbuildmat.2014.03.041.

[45] N. F. Salehuddin, M. B. Omar, R. Ibrahim, and K. Bingi, "A Neural Network-Based Model for Predicting Saybolt Color of Petroleum Products," *Sensors 2022, Vol. 22, Page 2796*, vol. 22, no. 7, p. 2796, Apr. 2022, doi: 10.3390/S22072796.

[46] A. Khan, H. Hwang, and H. S. Kim, "Synthetic Data Augmentation and Deep Learning for the Fault Diagnosis of Rotating Machines," *Mathematics*, vol. 9, no. 18, p. 2336, Sep. 2021, doi: 10.3390/math9182336.