2023, 20(4 Special Issue): 1451-1458
https://dx.doi.org/10.21123/bsj.2023.8504
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Deep Learning-based Predictive Model of mRNA Vaccine Deterioration: An Analysis of the Stanford COVID-19 mRNA Vaccine Dataset

Nisreen Sulayman* [1, 2]

[1]Department of Biomedical Engineering, Mechanical and Electrical Engineering Faculty, Damascus University, Damascus, Syria.
[2]Faculty of Informatics Engineering, Syrian Private University, Damascus, Syria.

Received 02/02/2023, Revised 04/08/2023, Accepted 06/08/2023, Published 30/08/2023

Abstract

The emergence of SARS-CoV-2, the virus responsible for the COVID-19 pandemic, has resulted in a global health crisis leading to widespread illness, death, and daily life disruptions. Having a vaccine for COVID-19 is crucial to controlling the spread of the virus which will help to end the pandemic and restore normalcy to society. Messenger RNA (mRNA) molecules vaccine has led the way as the swift vaccine candidate for COVID-19, but it faces key probable restrictions including spontaneous deterioration. To address mRNA degradation issues, Stanford University academics and the Eterna community sponsored a Kaggle competition.This study aims to build a deep learning (DL) model which will predict deterioration rates at each base of the mRNA molecule. A sequence DL model based on a bidirectional gated recurrent unit (GRU) is implemented. The model is applied to the Stanford COVID-19 mRNA vaccine dataset to predict the mRNA sequences deterioration by predicting five reactivity values for every base in the sequence, namely reactivity values, deterioration rates at high pH, at high temperature, at high pH with Magnesium, and at high temperature with Magnesium. The Stanford COVID-19 mRNA vaccine dataset is split into the training set, validation set, and test set. The bidirectional GRU model minimizes the mean column wise root mean squared error (MCRMSE) of deterioration rates at each base of the mRNA sequence molecule with a value of 0.32086 for the test set which outperformed the winning models with a margin of (0.02112). This study would help other researchers better understand how to forecast mRNA sequence molecule properties to develop a stable COVID-19 vaccine.

Keywords: COVID-19 Vaccine, Deep learning, mRNA, Predictive Model, Recurrent neural networks (RNN).

Introduction

The World Health Organization declared the novel coronavirus (COVID-19) a worldwide pandemic on 11 March 2020[1]. As of 10 August 2022, COVID-19 had caused over 584 million confirmed cases, including 6,418,958 deaths. Since then, COVID-19 has been the subject of extensive research[2-5]. To control the spread of the pandemic, governments worldwide have deployed considerable efforts and resources to develop COVID-19 vaccines[6]. Researchers have tested several approaches to COVID-19 vaccine development including mRNA[7,8].

Ribonucleic acid (RNA) is usually a single-stranded molecule. It has a backbone made of alternating phosphate groups and the sugar ribose, attached to each sugar is one of four bases: adenine (A), uracil (U), cytosine (C), and guanine (G). Various types of RNA exist in cells: mRNA, ribosomal RNA (rRNA),

2023, 20(4 Special Issue): 1451-1458
https://dx.doi.org/10.21123/bsj.2023.8504
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

and transfer RNA (tRNA)[9, 10]. While mRNA vaccines are relatively easier to manufacture compared to traditional inoculations, maintaining their stability remains a significant challenge. Researchers have found that mRNA vaccines are prone to degradation under various conditions such as changes in temperature or environment[11]. For example, a single cut can make the mRNA vaccine unusable[12]. The potential for damage to the mRNA vaccine is a significant concern, as it can render the vaccine ineffective. Currently, there is limited understanding of where the mRNA backbone is most vulnerable. Therefore, identifying stable mRNA COVID-19 vaccine molecules is essential for the widespread distribution and vaccination of large populations.

Ing et al[13], developed two machine learning algorithms to predict the deterioration rate of the mRNA vaccine for COVID-19. They used linear regression (LR) and light gradient boosting machines (LGBM) to build up deterioration prediction models using the Python language. Using a dataset comprising thousands of mRNA molecules with known deterioration rates at each position from the Eterna platform, the results showed that the LGBM performs better than linear regression (LR) when evaluated with the RMSE metric[13].

Muneer et al[14], looked into the potential of hybrid DL in predicting the deterioration of COVID-19 mRNA from mRNA sequences. Two deep hybrid neural network models were developed, one that combined graph convolutional neural networks (GCNs) and (GRU), and another that combined GCNs and convolutional neural networks (CNNs). The findings revealed that the GCN_GRU hybrid model was superior to the GCN_CNN model. Specifically, the GCN_GRU model had the highest MCRMSE and AUC scores[14].

Krishna et al[15] , clarified that the biggest challenge to mRNA vaccine production is structural instability. Using an appropriate sequence to vector representation, they applied three pre-trained gene embedding models (dna2vec, rna2vec, and lshvec) for predicting the amount of deterioration of the mRNA vaccine sequences. They compared the pre-trained models and found that dna2vec embedding performed best[15].

Imran et al[16], used a regularized long short-term memory network (LSTM) which is a type of recurrent neural network (RNN) to develop a prediction model to conclude the deterioration rates of the COVID-19 mRNA Vaccine. With the help of a Stanford dataset of the COVID-19 mRNA vaccine, they employed the LSTM to identify if and where mRNAs might be unstable under certain incubation measures. They found that LSTM performed better than tree-based algorithms using MCRMSE as a performance evaluation metric[16].

This study aims to use modern DL techniques, to design a prediction model for the deterioration of COVID-19 mRNA vaccine molecules to improve the stability of COVID-19 mRNA vaccines during the transportation process. The model will predict the probability of the deterioration rates for each position inside the mRNA sequence.

Materials and Methods

Dataset

On September, 11, 2020, researchers at Stanford University and the Eterna community partnered to sponsor a Kaggle competition to solve deterioration problems in mRNA. Researchers collected the data for 6034 mRNA sequences[17]. The dataset is available in the Kaggle repository, (https://www.kaggle.com/competitions/stanford-covid-vaccine/data).

The lengths of the mRNA sequences are 107 bases. mRNA sequences are available with labels containing the deterioration rates measured at different locations of the mRNA sequence, namely reactivity values (reactivity) and deterioration rates at high pH (deg pH10), high temperature (deg 50 C°), high pH with Magnesium (deg Mg pH10) and at high temperature with Magnesium (deg Mg 50 C°). For each mRNA sequence, three input features and five outputs\targets features were provided. Table 1 shows the input labels and their descriptions and examples used in this study. Output labels are listed in table 2. This study will develop a bidirectional GRU that takes the RNA sequence information as input and produces five predictions of deterioration rates at each base.

2023, 20(4 Special Issue): 1451-1458
https://dx.doi.org/10.21123/bsj.2023.8504
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Tables 1. Input labels used in this study and their descriptions and examples[17].

Feature	Sequence Length	Sequence type	Description	Example
Sequence	1x107 (Scored for 68 bases)	string	It represents the RNA sequence. It is a combination of four bases: adenine (A), uracil (U), cytosine (C), and guanine (G) for each sample.	GGAAAAGCUCUA…
Structure	1x107 (Scored for 68 bases)	string	It contains a sequence of '.', '(', and ')', indicating whether bases are paired or unpaired. Paired bases are denoted by opening and closing parentheses.	. . ..((((((.....)))).))..
Predicted loop type	1x107 (Scored for 68 bases)	string	It describes the structural context of each character in the sequence. They used bpRNA tool documented in [18]. It predicts the loop types as the following labels: hairpin loop (H), paired stem (S), multiloop (M), bulge (B), internal loop (I), external loop (X), and dangling end (E).	EEESSSHSSSSBSSX…

Table 2. Output (Reactivity) labels and their descriptions[17].

Feature	Sequence Length	Sequence type	Description
Reactivity	1x68	A vector of floating-point numbers	"It indicates the reactivity values for the first 68 bases in the sequence and is used to determine the likely secondary structure of the RNA sample".
Deg pH10	1x68	A vector of floating-point numbers	"It represents reactivity values for the first 68 bases as denoted in sequence. It is used to determine the likelihood of deterioration at the base/linkage after incubating without magnesium at high pH (pH 10)".
Deg Mg pH10	1x68	A vector of floating-point numbers	"It represents reactivity values for the first 68 bases as denoted in sequence and is used to determine the likelihood of deterioration at the base/linkage after incubating with magnesium at high pH (pH 10)".
Deg 50 C°	1x68	A vector of floating-point numbers	"It indicates reactivity values for the first 68 bases as denoted in sequence and is used to determine the likelihood of deterioration at the base/linkage after incubating without magnesium at a high temperature (50 degrees Celsius)".
Deg Mg 50 C°	1x68	A vector of floating-point numbers	"It indicates reactivity values for the first 68 bases as denoted in sequence and is used to determine the likelihood of deterioration at the base/linkage after incubating with magnesium at high temperature (50 degrees Celsius)".

Features engineering

Three categorical features are extracted from the mRNA, namely sequence (describes the RNA sequence), structure (indicates whether a base is paired or unpaired), and predicted loop type (defines the structural context). Sequence features result in four features by converting the A, U, G, and C base pair sequences to integer class vectors, structure features are extracted for each position of the mRNA resulting in three features, and the predicted loop results in seven features. Base method process for encoding the categorical features is applied[19].

Model Architecture

Sequence prediction is a dispute that requires using historical sequence information to predict the next values in the sequence. The proposed model relies on a bidirectional GRU which is a variation of traditional LSTM networks. Cho et al[20, 21], proposed the GRU to make each recurrent unit adaptively capture dependencies on a varied time scale. It

2023, 20(4 Special Issue): 1451-1458
https://dx.doi.org/10.21123/bsj.2023.8504
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

allows for the use of information from both directions (previous time steps and later time steps) to make predictions about the current state. In the bidirectional GRU model, categorical input features are encoded and fed into an embedding layer, which adds extra information to the neural network. In this study, the embedding layer extracted 200 features instead of the 14 input encoded features. The processed, encoded, and embedded features are then used as input for the bidirectional GRU sequence model layers.

The model consists of three layers of bidirectional GRUs, with 256 hidden units in each direction for each layer, resulting in 512 bidirectional units per layer. The bidirectional layers are utilized to improve the performance, by processing the data in both forward and backward directions to capture the information in the sequence data. The description of the bidirectional GRU model is depicted in table 3. The dropout process acts like a regularization parameter to overcome the overfitting problem in the learning process which makes the model less sensitive to the particular weights so that it can generalize better. Finally, the dense layer (Dense) has five outputs (reactivity values at each base) and the activation function is linear as it is a regression problem. Adam optimizer is used because of its advantages concerning memory, simplicity, and computation. The loss function is the MCRMSE as described in Eq1.

Table 3. Bidirectional GRU sequence model summary for layers, output shape, and number of parameters in each layer.

Layer (type)	Output shape	Number of parameters
input_4 (InputLayer)	[(None, 107, 3)]	0
embedding_3 (Embedding)	(None, 107, 3, 200)	2800
tf.reshape_2 (TFOpLambda)	(None, 107, 600)	0
spatial_dropout1d_2	(Spatial (None, 107, 600)	0
bidirectional_6	(Bidirection (None, 107, 512)	1317888
bidirectional_7	(Bidirection (None, 107, 512)	1182720
bidirectional_8	(Bidirection (None, 107, 512)	1182720
tf.__operators__.getitem_2	(None, 68, 512)	0
dense_2 (Dense)	(None, 68, 5)	2565
Total params: 3,688,693		
Trainable params: 3,688,693		
Non-trainable params: 0		

Evaluation metrics

The model's performance is evaluated using 68 bases of mRNA sequence, and the chosen evaluation metric is MCRMSE, which is represented by the following equation:

$$MCRMSE = \frac{1}{N_t}\sum_{j=1}^{N_t}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{ij}-\hat{y}_{ij})^2} \quad (1)$$

Here, N_t is the number of scored ground truth target columns, y is the ground truth value, and $\hat{y}$ is the predicted value. Five deterioration rate parameters to be predicted namely: reactivity, deg_pH10, deg_Mg_pH10, deg_50 C°, and deg_Mg_50 C°. Lower MCRMSE indicates a better prediction of the deterioration rates.

Model Training

The model is applied to the Stanford COVID-19 mRNA vaccine dataset[17], the dataset of this study encompassed 2400 instances. It is split into the training set which contains 1944 samples for the construction of the bidirectional GRU model, the validation set which contains 240 samples for the tuning of the model, and the test set which contains 216 samples to measure the generalization performance. Table 4 and Table 5 show the most important configurations of the model and the training options respectively.

2023, 20(4 Special Issue): 1451-1458
https://dx.doi.org/10.21123/bsj.2023.8504
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 4. Bidirectional GRU configurations.

Configuration	Value
Sequence length (seq_len)	107
Prediction length (pred_len)	68
Dropout	0.5
embed_dim	200
hidden_dim	256

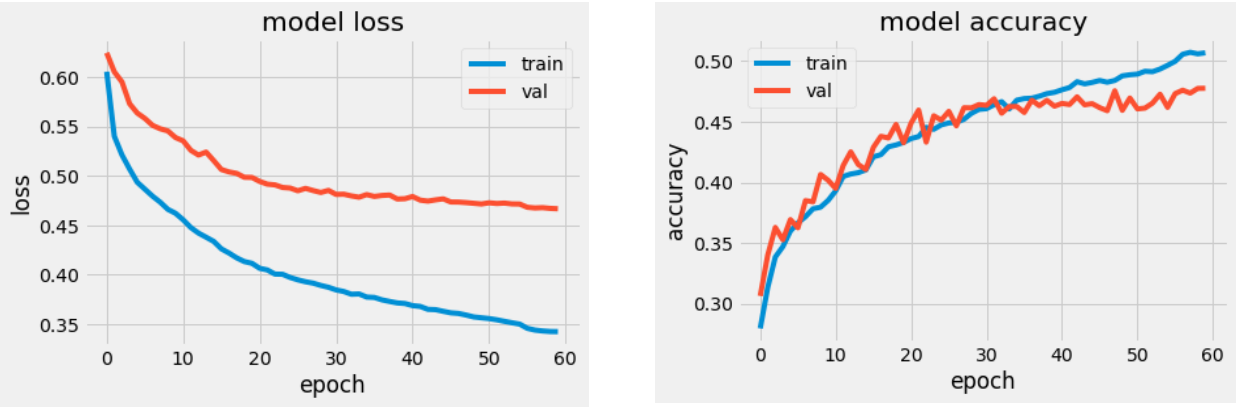
Table 5. Bidirectional GRU training options.

Property	Value
Batch size	64
Epochs	60
Verbose	2
Optimizer	Adam
Learning rate	0.001
Reduce learning rate on a plateau	patience=5

Results and Discussion

The bidirectional GRU model predicts five target columns of reactivity values at each base of the mRNA sequence, namely: reactivity, deg Mg_pH10, deg_pH10, deg_Mg_50 C°, and deg_50 C°. The performance evaluation metrics are the MCRMSE as is given in equation 1 and the accuracy.

As shown in Fig 1 the training and validation curves converged steadily with an increasing number of epochs. Table 6 summarizes the loss and accuracy values achieved by the bidirectional GRU model on the training, validation, and test datasets respectively.



(a) **(b)**

Figure 1. Bidirectional GRU performance evaluation curves. (a) Training and validation loss curves. (b) Training and validation accuracy curves.

Table 6. Bidirectional GRU performance evaluation metrics: loss and accuracy.

	Training data	Validation data	Test data
Loss (MCRMSE)	0.32536	0.46672	0.32086
Accuracy	0.54063	0.47751	0.49986

I chose the bidirectional GRU because it is less complex and faster to compute when compared to other RNNs[20,21]. The mRNA molecules dataset used in this study is derived from a Kaggle competition launched by Stanford researchers. MCRMSE is used as a performance metric to score the competition's winners.

The prize winner's performing model scores were: 0.34198, 0.34266, and 0.34327 respectively based on MCRMSE[17]. The difference from 1st to 3rd place is marginal (0.00129). Results in this study (MCRMSE=0.32086) outperformed the winning models with a marginal of (0.02112).

On the other hand, when comparing it to the literature, I chose related works considering the same benchmark dataset (Stanford dataset on Kaggle) and evaluation metric (MCRMSE). Imran et al[16], used a regularized LSTM network to conclude the deterioration rates of COVID-19 mRNA Vaccine. They found that LSTM performed better than tree-

2023, 20(4 Special Issue): 1451-1458
https://dx.doi.org/10.21123/bsj.2023.8504
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

based algorithms with MCRMSE of 0.4904 and 0.5165 for the training and validation sets compared to 0.32536 and 0.46672 in this study on the training and validation sets[16]. Ing et al[13], used two machine learning algorithms to predict the deterioration rate of mRNA vaccine for COVID-19. They utilized linear regression (LR) and an LGBM. The results show that (LGBM) outperformed linear regression with an RMSE value of 0.22465 compared to 0.39574 for linear regression.

Conclusion

In this study, five deterioration reactivity values for every position at each base in the mRNA sequence were predicted. The bidirectional GRU uses the Stanford dataset to build a deterioration prediction for the COVID-19 mRNA vaccine model. Although deterioration prediction models prove to be helpful in finding stable mRNA vaccine candidates, there are several limitations that need to be addressed. A significant limitation of this study is the length of the mRNA sequences used. The length of the mRNA molecule used in this study is 107 bases and the deterioration rates were scored for the first 68 bases of the sequence, while an actual COVID-19 mRNA vaccine would likely be longer[22], which indicates further research exploring the reliability of such algorithms in predicting longer sequences is encouraged.

In the future, various features based on the characteristics of mRNA sequences could be used to enhance the accuracy of the models, and more DL models for sequence predictions could be applied. Other considerations to improve the results include but are not limited to be using pre-trained models for deterioration predictions, augmentation techniques to increase the size of the dataset, and utilizing cross-validation on the DL algorithms. It is hoped that this study will be of some use to other researchers in creating a better understanding of prediction models for mRNA sequence molecules.

Acknowledgment

The author would like to thank Damascus University for providing the necessary approval and support, which made this research possible.

Authors' Declaration

- Conflicts of Interest: None.
- I hereby confirm that all the Figures and Tables in the manuscript are mine. Furthermore, any Figures and images, that are not mine, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- Author sign on ethical consideration's approval.
- Ethical Clearance: The project was approved by the local ethical committee in university of Damascus, Mechanical and Electrical Engineering Faculty, Department of Biomedical Engineering on 24/7/2022.

References

1. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. Acta Bio Medica: Atenei Parmensis. 2020; 91(1): 157. https://dx.doi.org/10.23750/abm.v91i1.9397
2. Sulayman N. Comparative study for automated coronavirus detection in CT images with transfer learning. Damascus Uni J Eng Sci, 2022; 38(5): 245-255. http://journal.damascusuniversity.edu.sy/index.php/engj/article/download/2248/1775
3. Mohammed SK, Taha MM, Taha EM, Mohammad MN. Cluster analysis of biochemical markers as predictor of COVID-19 severity. Baghdad Sci J. 2022; 19(6 (Suppl.)): 1423-1429 https://dx.doi.org/10.21123/bsj.2022.4242
4. Zaki SM, Jaber MM, Kashmoola MA. Diagnosing COVID-19 Infection in Chest X-Ray Images Using Neural Network. Baghdad Sci J. 2022; 19(6): 1356-1361. http://dx.doi.org/10.21123/bsj.2022.5965
5. Mohammed TS, Sultan AI, Alheeti KM, Aljebory KM, Sultan HI, Al-Ani MS. COVID-19 Diagnosis Using Spectral and Statistical Analysis of Cough Recordings Based on the Combination of SVD and DWT. Baghdad Sci J. 2022; 0372-. https://doi.org/10.21123/bsj.2022.6516

2023, 20(4 Special Issue): 1451-1458
https://dx.doi.org/10.21123/bsj.2023.8504
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

6. WHO Coronavirus (COVID-19) Dashboard, https://covid19.who.int/ (accessed on 11 August 2022)
7. Corbett KS, Edwards DK, Leist SR, Abiona OM, Boyoglu-Barnum S, Gillespie RA, Himansu S, Schäfer A, Ziwawo CT, DiPiazza AT, Dinnon KH. SARS-CoV-2 mRNA vaccine design enabled by prototype pathogen preparedness. Nature. 2020; 586(7830): 567-71. https://doi.org/10.1038/s41586-020-2622-0
8. Kremsner PG, Mann P, Kroidl A, Leroux-Roels I, Schindler C, Gabor JJ, et al. Safety and immunogenicity of an mRNA-lipid nanoparticle vaccine candidate against SARS-CoV-2: A phase 1 randomized clinical trial. Wien Klin Wochenschr. 2021; 133(17-18): 931-41. https://doi.org/10.1007/s00508-021-01922-y
9. Ribonucleic Acid (RNA). https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid (Accessed on 11 August 2022)
10. Lyngsø RB, Pedersen CN. RNA pseudoknot prediction in energy-based models. J Comput Biol. 2000; 7(3-4): 409-27. https://doi.org/10.1089/106652700750050862
11. Pardi N, Hogan MJ, Porter FW, Weissman D. mRNA vaccines—a new era in vaccinology. Nat Rev Drug Discov. 2018; 17(4): 261-79. https://doi.org/10.1038/nrd.2017.243
12 Bouarara HA. N-Gram-Codon and Recurrent Neural Network (RNN) to Update Pfizer-BioNTech mRNA Vaccine. Int J Comput Intell Syst. 2022; 14(1): 1-24. http://doi.org/10.4018/IJSSCI.305838
13. Ing SH, Abdullah AA, Harun NH, Kanaya S. COVID-19 mRNA Vaccine Degradation Prediction Using LR and LGBM Algorithms. J Phys Conf Ser. 2021; 1997(1): 012005. IOP Publishing. http://doi.org/10.1088/1742-6596/1997/1/012005
14 Muneer A, Fati SM, Akbar NA, Agustriawan D, Wahyudi ST. iVaccine-Deep: Prediction of COVID-19 mRNA vaccine degradation using deep learning. J King Saud Univ- Comput. Inf Sci. 2022; 34(9): 7419-32. https://doi.org/10.1016/j.jksuci.2021.10.001
15. Krishna UV, Premjith B, Soman KP. A Comparative Study of Pre-trained Gene Embeddings for COVID-19 mRNA Vaccine Degradation Prediction. Proc 7th Int Conf Math Comput : pp. 301-308. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-6890-6_22
16. Imran SA, Islam MT, Shahnaz C, Islam MT, Imam OT, Haque M. COVID-19 mRNA Vaccine Degradation Prediction using Regularized LSTM Model. Int Women Eng Conf. 2020; (pp. 328-331). IEEE. https://doi.org/10.1109/WIECON-ECE52138.2020.9398044
17. Openvaccine: Covid-19 mrna vaccine degradation prediction, Sep. 2020; https://www.kaggle.com/competitions/stanford-covid-vaccine/data
18. Danaee P, Rouches M, Wiley M, Deng D, Huang L, Hendrix D. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. Nucleic Acids Res. 2018; 46(11): 5381-94. https://doi.org/10.1093/nar/gky285
19. Qaid TS, Mazaar H, Alqahtani MS, Raweh AA, Alakwaa W. Deep sequence modelling for predicting COVID-19 mRNA vaccine degradation. Peer J Comput Sci. 2021; 7: e597. https://doi.org10.7717/peerj-cs.597
20. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proc Conf Empir Methods Nat Lang Proc. 2014; 1406.1078. https://doi.org/10.48550/arXiv.1406.1078
21. Cho K, Van Merriënboer B, Bahdanau D, Bengio Y. On the properties of neural machine translation: Encoder-decoder approaches. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. 2014; 1409.1259. https://doi.org/10.48550/arXiv.1409.1259
22. Schoenmaker L, Witzigmann D, Kulkarni JA, Verbeke R, Kersten G, Jiskoot W, et al. mRNA-lipid nanoparticle COVID-19 vaccines: Structure and stability. Int J Pharm X. 2021; 601:120586. https://doi.org/10.1016/j.ijpharm.2021.120586

2023, 20(4 Special Issue): 1451-1458
https://dx.doi.org/10.21123/bsj.2023.8504
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

نموذج تعلّم عميق تنبّؤي بتلف لقاح الرنا المرسال (mRNA): تحليل قاعدة بيانات ستانفورد للقاح الرنا المرسال المضاد لكوفيد-19

نسرين سليمان[1،2]

[1]قسم الهندسة الطبية، كلية الهندسة الميكانيكية والكهربائية، جامعة دمشق، دمشق، الجمهورية العربية السورية
[2]قسم الذكاء الصنعي وعلوم البيانات، كلية الهندسة المعلوماتية، الجامعة السورية الخاصة، دمشق، الجمهورية العربية السورية

الخلاصة

أدى ظهور فيروس SARS-CoV-2، الفيروس المسؤول عن جائحة كوفيد-19، إلى أزمة صحية عالمية أدت إلى انتشار المرض والوفاة واضطرابات واسعة النطاق في الحياة اليومية. يعد الحصول على لقاح لجائحة كوفيد-19أمراً بالغ الأهمية في السيطرة على انتشار الفيروس مما سيساعد على إنهاء الوباء واستعادة الحياة الطبيعية للمجتمع. يعد لقاح الرنا المرسال (mRNA) من اللقاحات الأكثر حظاً للاستخدام كلقاح مضاد لفيروس كوفيد-19، إلا أنه يواجه عدد من قيود الاستخدام المرتبطة بالتلف التلقائي. قام أكاديميون من جامعة ستانفورد ومجتمع Eterna برعاية مسابقة عبر منصة Kaggle لدراسة مسألة التلف التلقائي لجزيء الرنا المرسال. يهدف هذا البحث إلى بناء نموذج تعلّم عميق تنبّؤي بمعدل التلف الحاصل عند كل جزيء من جزئيات لقاح الرنا المرسال. تم بناء نموذج تعلم عميق باستخدام الشبكات العصبونية التكرارية ثنائية الاتجاه واستخدام هذا النموذج على قاعدة بيانات الرنا المرسال التي وفرتها جامعة ستانفورد على الانترنت كلقاح مضاد للكوفيد-19 للتنبؤ بمعدل التلف الحاصل لتسلسل الرنا المرسال. فقد تم التنبؤ بخمسة قيم فعالية عند كل جزيء في تسلسل الرنا المرسال. شملت القيم التي تم التنبؤ بها معدل التلف التلقائي عند: ارتفاع درجة الحموضة، ارتفاع درجة الحرارة، ارتفاع درجة الحموضة بوجود المغنيسيوم وارتفاع درجة الحرارة بوجود المغنيسيوم. تمت تجزئة قاعدة بيانات ستانفورد للقاح الرنا المرسال المضاد لكوفيد-19 إلى ثلاث مجموعات رئيسة تشمل: التدريب والتحقق ومجموعة الاختبار. بلغت قيمة متوسط جذر الخطأ التربيعي لقيم معدل التلف التلقائي عند كل جزيء من تتابع جزئيات لقاح الرنا المرسال 0.32086 في مجموعة الاختبار. وقد تفوقت هذه القيمة على النماذج الفائزة في المسابقة بهامش0.02112. تساعد هذه الدراسة الباحثين الآخرين على فهم أفضل لكيفية التنبؤ بخصائص جزيء تسلسل الرنا المرسال لتطوير لقاح مستقر كوفيد-19.

الكلمات المفتاحية: لقاح كوفيد-19، التعلم العميق، الرنا المرسال، نموذج تنبؤي، الشبكات العصبونية التكرارية.