SPIKE PROTEIN MUTATION ANALYSIS OVER VARIANTS USING DEEP LEARNING

A MAJOR PROJECT REPORT

Submitted by

CH.EN.U4AIE20001	ANBAZHAGAN E
CH.EN.U4AIE20053	RAMYA POLAKI
CH.EN.U4AIE20069	TRINAYA KODAVATI
CH.EN.U4AIE20031	SMITHIN REDDY K
CH.EN.U4AIE20035	ABHIRAM KUNCHAPU

in partial fulfilment for the award of the degree

Of

BACHELOR OF TECHNOLOGY

IN

INTELLIGENCE OF BIOLOGICAL SYSTEMS

AMRITA SCHOOL OF ENGINEERING, CHENNAI

AMRITA VISHWA VIDYAPEETHAM

CHENNAI – 601103, TAMIL NADU

AMRITA VISHWA VIDYAPEETHAM

AMRITA SCHOOL OF ENGINEERING, CHENNAI, 601103

BONAFIDE CERTIFICATE

This is to certify that the major project report entitled **“SPIKE PROTEIN MUTATION ANALYSIS OVER VARIANTS USING DEEP LEARNING”** submitted by

CH.EN.U4AIE20001	ANBAZHAGAN E
CH.EN.U4AIE20053	RAMYA POLAKI
CH.EN.U4AIE20069	TRINAYA KODAVATI
CH.EN.U4AIE20031	SMITHIN REDDY K
CH.EN.U4AIE20035	ABHIRAM KUNCHAPU

in partial fulfillment of the requirements for the award of the **bachelor of Master of Technology** in **COMPUTER SCIENCE ENGINEERING** is a bonafide record of the work carried out under my guidance and supervision at Amrita School of Engineering, Chennai.

Signature

Dr.I.R.Oviya

CSE Dept Assistant Professor

This project report was evaluated by us on ……………...

INTERNAL EXAMINER EXTERNAL EXAMINER

ACKNOWLEDGEMENT:

I offer my sincere pranams at the lotus feet of Universal guru, MATA AMRITANANDAMAYI DEVI who blessed me with her grace to make this a successful project.

I express my deep sense of gratitude to Dr. P Shankar, Principal, Amrita School of Engineering, Chennai for his kind support. I would like to extend my gratitude and heartfelt thanks to Dr. I.R.Oviya, CSE Department, for his constant help, suggestions, and inspiring guidance. I am grateful to my guide, the Department of CSE, for his invaluable support and guidance during the course of the major project work. I am also indebted to Dr. Prasanna Kumar, Chairperson of the Department of AI, ASE, Chennai for his guidance.

I especially thank our director Sri I B Manikandan who supported us in every possible manner.

INDEX

Topic **Page no.**

1. ABSTRACT:

The ability to predict pathogen behavior will increase disease control, prevention, and treatment greatly. Despite tremendous progress in other areas, deep learning has yet to make a contribution to the challenge of predicting mutations in evolving populations. To close this gap, we'll create a novel machine learning framework that combines genomic adversarial networks (GANs) and recurrent neural networks (RNNs) to reliably anticipate gene mutations and biological population dynamics in the future. The inverted and extended phylogenetic model of protein evolution was used. We'll use MutaGAN, an adversarial framework, to train a sequencer to build entire protein sequences augmented with future viral population mutations. Different versions of the SARSCoV2 viral sequence will be employed. In this paper, we propose an idea to predict the mutation of spike protein using an amino acid dataset with the help of RNN and GAN.

2. INTRODUCTION

2.1 Spike Protein Mutation:

With the delta COVID-19 variation, also known as B.1.617.2, rapidly spreading over the world, we need to figure out what makes this variety more contagious. According to one study, the delta version is 60% more transmissible than the alpha variant. The delta variant of SARS-CoV-2, B.1.617.2, has 23 mutations compared to the original COVID-19 strain (alpha strain). Twelve of these mutations can be found in the spike protein.

The spike protein induces host cells to stick together, allowing the virus to enter. The immune system recognizes the spike protein as a target for viral eradication. B cells produce antibodies to bind to and destroy the spike protein when the immune system recognizes it as foreign.

The spike protein is made up of two subunits, S1 and S2. S1 binds to the ACE2 receptor, while S2 aids in the virus's fusion and integration with the host cell. The immune system's ability to recognize spike proteins and bind antibodies to them becomes more difficult as they evolve, making it more difficult to destroy the virus.

2.2 Functional Variation:

When compared to the Alpha variant, the Delta variant is linked to higher infectious virus loads and lower antiviral IgG levels in the upper respiratory tract.

This suggests that Delta's fitness has improved, maybe as a result of viral genetic factors that improve infectivity or replication. Changes in the Delta variant's spike protein are thought to increase the S protein's cleavage efficiency, allowing for better binding to the host cell receptor (ACE2). The S: P681R change, in particular, may increase the S protein's cleavage efficiency, allowing for better binding to the host cell receptor (ACE2). It's licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives-NoDerivatives-NoDerivatives-NoDerivatives-NoDerivatives-NoDe (which was not peer-reviewed) is the author/funder, who has given medRxiv perpetual permission to exhibit the preprint.

In one of the researches, the spike mutation P681R has been discovered as a crucial predictor of the Delta variant's improved viral replication fitness when compared to the Alpha form. Because of the enhanced furin cleavage site, the P681R mutation improves spike protein digestion. Spike mutations that impact furin cleavage efficiency, as well as other changes that may increase viral replication, pathogenicity, and/or immune evasion, must be continuously studied when new variations appear.

2.3 MutaGAN:

The basic function of GAN is to train a generator and discriminator in an adversarial way. GAN's primary job is to train an adversarial generator and discriminator. LSTM networks are a sort of recurrent neural network that can learn order dependence in sequence prediction issues. This is necessary in complicated problem fields such as machine translation, speech recognition, and others.

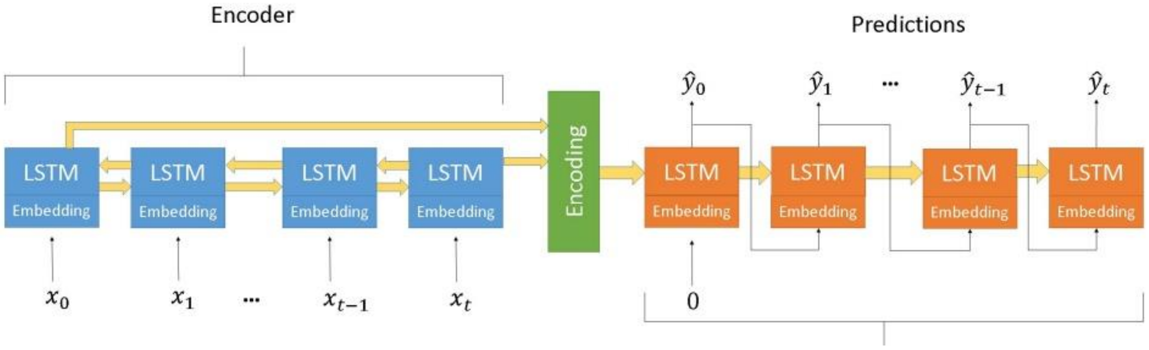


Figure 1: A seq2seq model using two bidirectional LSTMs encoder and a unidirectional LSTM decoder and embedding layer [15]

A seq2seq translation deep neural network with LSTMs and embedding layers generates the MutaGAN. A bidirectional LSTM is used in the encoding layer. The encoder's output is mixed with a vector of random noise generated by a normal distribution N. (0,1). A softmax dense layer receives the output of the LSTM decoder. Instead of using a probability distribution, the argmax function is used to select a single amino acid at each position. The discriminator employs a somewhat different structure than the generator's encoder, although it employs the same weights. This is due to the fact that an argmax function cannot be differentiated. As a result, the discriminator's initial encoder layer is a linear dense layer with the same output size as the generator's term embedding layer. This enables it to accept the dense layer decoder output from the generator as input. This dense layer has the same weights as the embedding layer, resulting in a linear combination of the embeddings from the encoder's embedding layer. The discriminator accepts two sequences as input and assesses whether they are a true parent-child pair or not. A parent and generated sequences, as well as two real sequences that are not parent-child pairings, are the sequences that are not real parent-child pairs. [15]

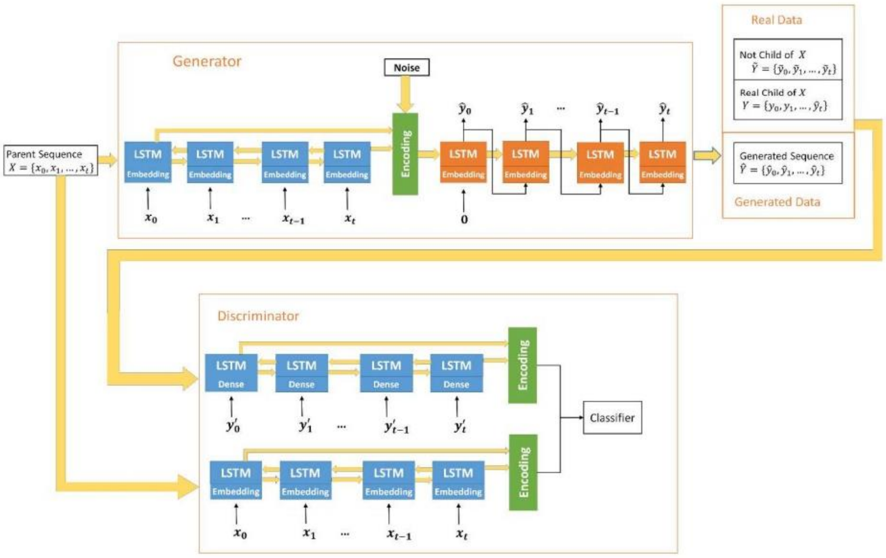


Figure 2: The MutaGAN framework's architecture [15]

2.4 Phylogenetic (tree) analysis:

Branching diagrams are used in phylogenetic analysis to depict the evolutionary history or relationship between distinct species, animals, or properties of an organism (genes, proteins, organs, and so on) that developed from a common ancestor.

Phylogenetic analysis is useful for discovering biological diversity, genetic classifications, and evolutionary developmental processes.

3. LITERATURE REVIEW

The capacity to foresee a pathogen's development would greatly enhance disease management, prevention, and treatment. Despite making substantial progress in other areas, deep learning has failed to make a major contribution to the problem of forecasting mutations in evolving populations. To fill this gap, in the paper, they created a new machine learning framework that combines generative adversarial networks (GANs) and recurrent neural networks (RNNs) to properly anticipate genetic mutations and future biological population dynamics. Researchers have trained a sequence-to-sequence generator within an adversarial framework called MutaGAN to generate complete protein sequences augmented with possible mutations of future virus populations using a generalized time-reversible phylogenetic model of protein evolution with bootstrapped maximum likelihood tree estimation. Because influenza virus sequences are a significant human pathogen with new strains emerging every year, and global surveillance efforts have generated a large amount of publicly available data from the National Center for Biotechnology Information's (NCBI) Influenza Virus

Resource, they were chosen as an ideal test case for this deep learning framework (IVR). MutaGAN generated "child" sequences with a median Levenshtein distance of 2.00 amino acids from a given "parent" protein sequence. In addition, the generator was able to add at least one mutation from the worldwide influenza virus population to the majority of parent proteins. These findings show how powerful the MutaGAN framework is for pathogen forecasting, with implications for any protein population's evolutionary prediction.

In recent research, the viability of antiviral medicines is nevertheless hampered by viral progress. +e capacity to predict this evolution will aid in the early diagnosis of drug-resistant strains and may promote antiviral medications to be the most effective treatment option. A deep learning model known as the seq2seq neural network has evolved in recent years and is commonly utilized in natural language processing. This method is used in this study to predict next-generation sequences using the seq2seq LSTM neural network while treating the sequences as text data. They employed hot single vectors as input to the model, and it keeps the fundamental information location of each nucleotide in the sequences as a result. The suggested model was evaluated using two RNA virus sequencing datasets, with promising results. +e the results show how the LSTM neural network for DNA and RNA sequences may be used to solve various sequencing problems in bioinformatics.

In research, the depict RNA virus has the capability of causing mutations in humans. To understand the development of this virus and assess the danger of emerging infectious illness, precise mutation rates must be determined. The study investigates the mutation rate of the whole genome sequence based on patient data from various nations. The obtained data is analyzed individually to determine nucleotide and codon mutations. Furthermore, the calculated mutation rate is divided into four categories depending on the size of the dataset: China, Australia, the United States, and the rest of the world. Thymine (T) and Adenine (A) are altered to different nucleotides in large amounts throughout all areas, however, codons do not mutate as frequently as nucleotides.

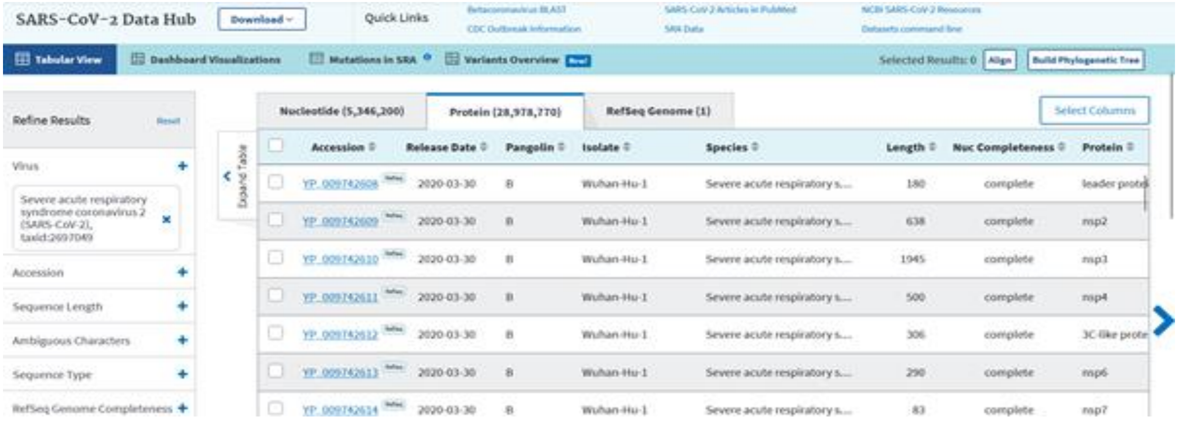
The future mutation rate of this virus was predicted using a recurrent neural network-based Long ShortTerm Memory (LSTM) model. The LSTM model produces an optimized Root Mean Square Error (RMSE) of 0.06 in testing and 0.04 in training. The nucleotide mutation rate of the 400th patient in the future has been anticipated using this training and testing method. Mutating nucleotides from T to C and G, C to G, and G to T results in a 0.1 percent increase in mutation rate. While changing T to A and A to C results in a 0.1 percent decrease. If more patient data is available in updated time, this model may be used to estimate day basis mutation rates.

The lethal coronavirus disease 2019 (COVID-19) pandemic, which is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has spread worldwide. Despite many efforts by scientists, medical experts, healthcare professionals, and society as a whole, the slow progress in drug discovery and antibody therapeutic development, the unknown potential side effects of existing vaccines, and the high transmission rate of SARS-CoV-2 remind us of the sad reality that our current understanding of SARS-CoV-2 transmission, infectivity, and evolution is unfortunately very limited.

4. METHODOLOGY

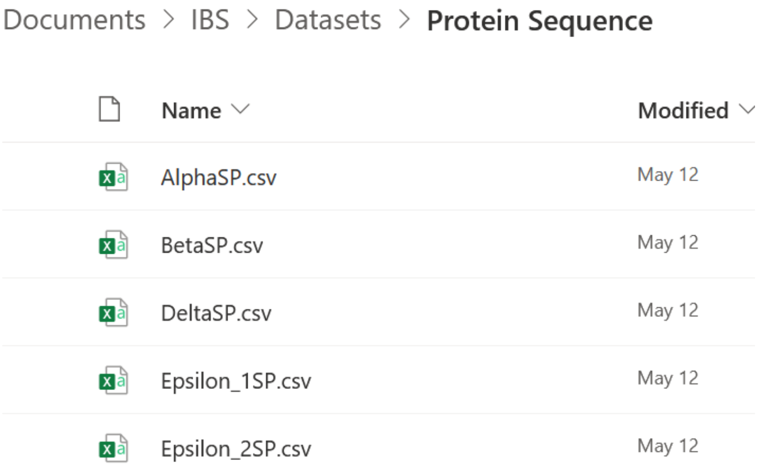
4.1 Data Collection:

We had to collect and generate our own dataset as the first step of our model because there was none available. As a result, the NCBI Virus Database assisted us in getting Covid spike protein Fasta Sequences in terms of Accession ID List (Fig. 4.1) to.csv data (Fig. 4.2). Using the Accession-ID List [5], we had to retrieve the Sequences directly from NCBI. We were able to manifest this extraction of fasta sequences using SeqIO API [6] thanks to the BioPython package in Python.



Accession	Release Date	Pangolin	Isolate	Species	Length	Nuc Completeness	Protein
YP_009742608	2020-03-30	B	Wuhan-Hu-1	Severe acute respiratory s...	180	complete	leader prote
YP_009742609	2020-03-30	B	Wuhan-Hu-1	Severe acute respiratory s...	638	complete	nsp2
YP_009742610	2020-03-30	B	Wuhan-Hu-1	Severe acute respiratory s...	1945	complete	nsp3
YP_009742611	2020-03-30	B	Wuhan-Hu-1	Severe acute respiratory s...	500	complete	nsp4
YP_009742612	2020-03-30	B	Wuhan-Hu-1	Severe acute respiratory s...	306	complete	3C-like prote
YP_009742613	2020-03-30	B	Wuhan-Hu-1	Severe acute respiratory s...	290	complete	nsp6
YP_009742614	2020-03-30	B	Wuhan-Hu-1	Severe acute respiratory s...	83	complete	nsp7

Figure 3: NCBI DataHub


Documents > IBS > Datasets > Protein Sequence

Name	Modified
AlphaSP.csv	May 12
BetaSP.csv	May 12
DeltaSP.csv	May 12
Epsilon_1SP.csv	May 12
Epsilon_2SP.csv	May 12

Figure 4: Accession ID List of Variants

Nextstrain Clade	Pango Lineage	WHO Label
Alpha	B.1.17	Alpha α
Beta	B.1.351	Beta β
Gamma	P.1	Gamma γ
Delta	B.1.617.2	Delta δ
Kappa	B.1.617.1	Kappa κ
Epsilon	B.1.427,B.1.429	Epsilon ε
Eta	B.1.525	Eta η
Iota	B.1.526	Iota ι
Lambda	C.37	Lambda λ
Mu	B.1.621	Mu μ
Omicron	BA.1	Omicron o

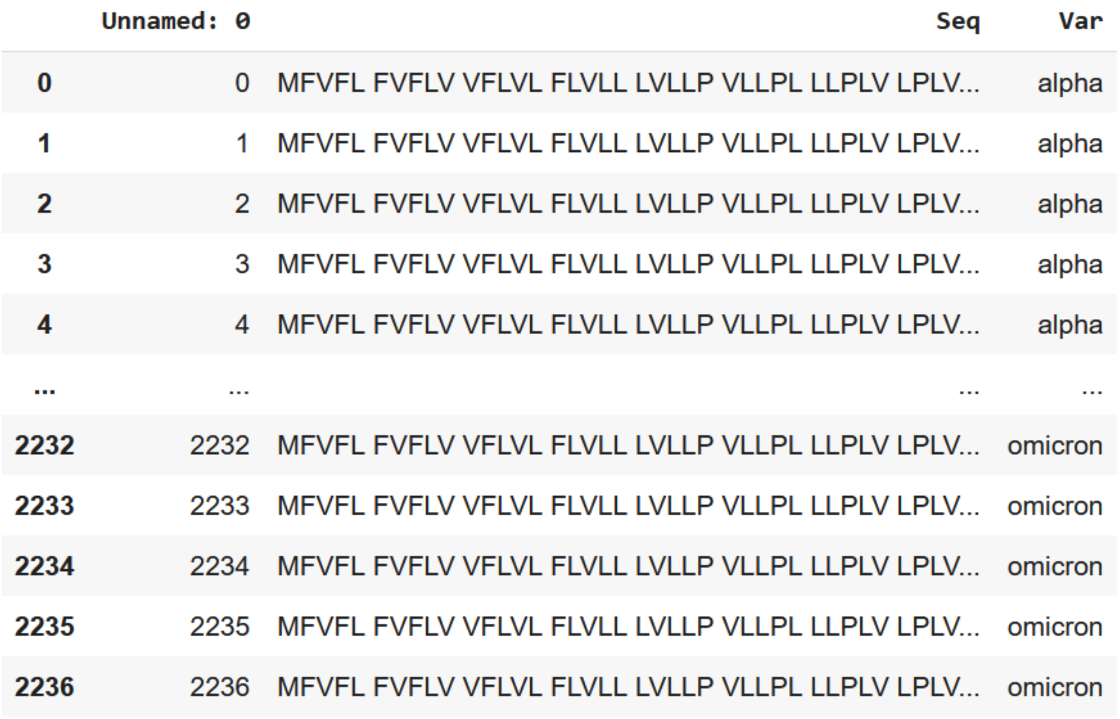
Table. 4. 1 Variants and its Panglin [7]

To move forward, we must first understand Pango Lineage (PangoLin). PANGOLIN (Phylogenetic Assignment of Named Global Outbreak Lineages) is a software tool created by Andrew Rambaut's team, with an online application developed by the Centre for Genomic Pathogen Surveillance in South Cambridgeshire. [7] Its purpose is to apply the PANGO nomenclature to classify SARS-CoV-2, the virus that causes COVID-19, into genetic lineages. A user can submit a SARS-CoV-2 sample's entire genome sequence, which is then compared to other genome sequences and the most likely lineage assigned (PANGO lineage). [7] According to Pangolin in Table .4.1 [7] , We can notice that every distinct variant is provided with unique Pango Lineage. NCBI has a Filter to segregate the variants with Pangolin[5].

4.2 Data Pre-processing:

Before training the model, we must convert the sequences in the form of phrases into k-mers, with k preferably greater than 5. K-mers are sub-sequences in a sequence having a specific size of k. So fragmenting them in order , will represent the sequence as a sentences with len(sequence)-(k-1) words as in Fig 4.4 and also, it's been done to all the variant datasets and stored.

SEQUENCE 1: MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...



	Unnamed: 0	Seq	Var
0	0	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	alpha
1	1	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	alpha
2	2	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	alpha
3	3	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	alpha
4	4	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	alpha
...	...	...	...
2232	2232	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	omicron
2233	2233	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	omicron
2234	2234	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	omicron
2235	2235	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	omicron
2236	2236	MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLV...	omicron

Fig 4. 4 Fragmentation of Sequences

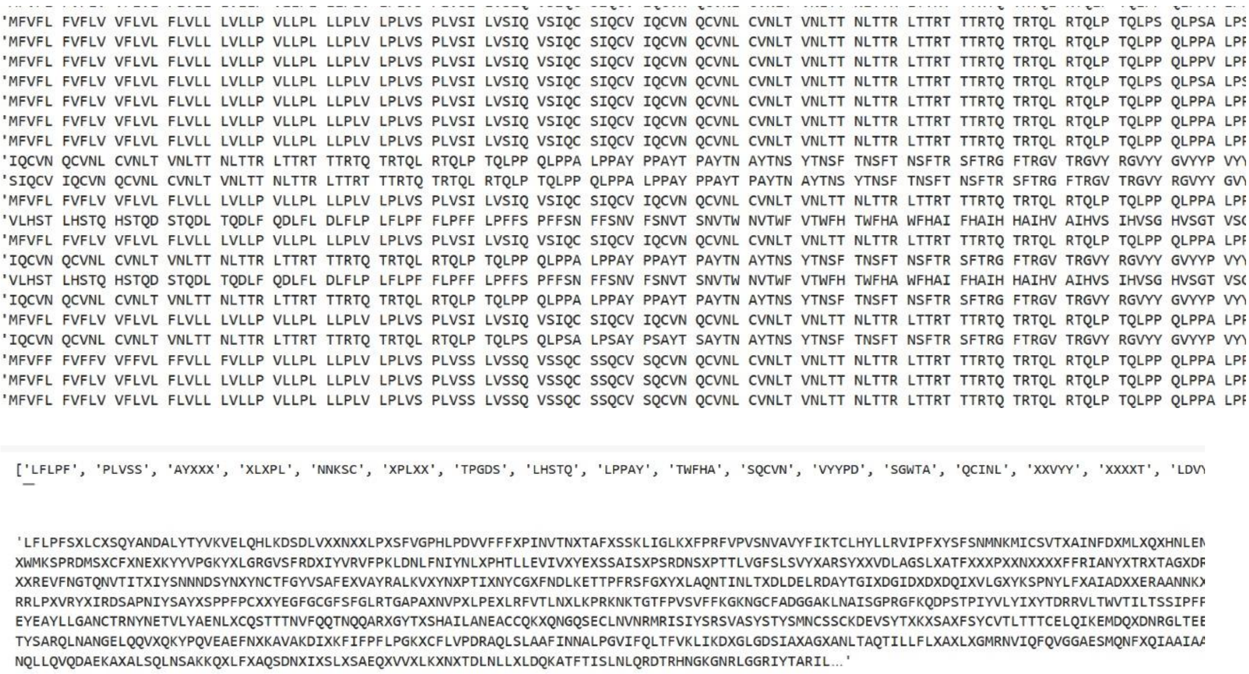
One Hot Encoding & Padding:

Every word (including symbols) in the specified text input is encoded as a vector of just 1 and 0 values in one-hot encoding. As a result, a one-hot vector consists of only one and zero elements. Each word is encoded or written as a single-hot vector. This allows the one-hot vector to uniquely identify the word and vice versa, ensuring that no two words are represented by the same one-hot vector. There are ample tools to work with in Keras for Natural Language Processing. One of the functions in the preprocessing module for text modification is the "one hot" function. This function takes a string of text as input and outputs a list of encoded integers, each of which corresponds to a word (or token) in the given string.After encoding and adding, x values are scaled to get high accuracy while training

5. DISCUSSION AND RESULT

After training the model, we have achieved a feasible accuracy MutaGAN we have pulled out the fake predictions that our generator generated and discriminator outputted. That could be possibly new mutations to occur. It will result in new fragments, so that time it can be a new variant. Another is we can use an algorithm to get the next suggestion of fragments to the prior fragments and generate new sequences and comparing it with another input sequence might tell us about its variant type. SARS cov2 spike protein, also called surface glycoprotein, is a virulent sequence that enters human host cells to affect the genome.

In our Research, we trained a model to give a false sequence generated from its parent sequenced and analyzed.



```
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPS QLPSA LPS
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPV LPF
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPS QLPSA LPS
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
'IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPPAY PPAYT PAYTN AYTNS YTNSF TNSFT NSFTR SFTRG FTRGV TRGVY RGVYY GVYYP VYY
'SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPPAY PPAYT PAYTN AYTNS YTNSF TNSFT NSFTR SFTRG FTRGV TRGVY RGVYY GVY
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
'VLHST LHSTQ HSTQD STQDL TQDLF QDLFL DLFLP LFLPF FLPFF LPFFS PFFSN FFSNV FSNVT SNVTW NVTWF VTWFH TWFHA WFHAI FHAIH HAIHV AIHVS IHVSG HVSGT VSC
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
'IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPPAY PPAYT PAYTN AYTNS YTNSF TNSFT NSFTR SFTRG FTRGV TRGVY RGVYY GVYYP VYY
'VLHST LHSTQ HSTQD STQDL TQDLF QDLFL DLFLP LFLPF FLPFF LPFFS PFFSN FFSNV FSNVT SNVTW NVTWF VTWFH TWFHA WFHAI FHAIH HAIHV AIHVS IHVSG HVSGT VSC
'IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPPAY PPAYT PAYTN AYTNS YTNSF TNSFT NSFTR SFTRG FTRGV TRGVY RGVYY GVYYP VYY
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSI LVSIQ VSIQC SIQCV IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
'IQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPS QLPSA LPSAY PSAYT SAYTN AYTNS YTNSF TNSFT NSFTR SFTRG FTRGV TRGVY RGVYY GVYYP VYY
'MFVFF FVFFV VFFVL FFVLL FVLLP VLLPL LLPLV LPLVS PLVSS LVSSQ VSSQC SSQCV SQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSS LVSSQ VSSQC SSQCV SQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
'MFVFL FVFLV VFLVL FLVLL LVLLP VLLPL LLPLV LPLVS PLVSS LVSSQ VSSQC SSQCV SQCVN QCVNL CVNLT VNLTT NLTTR LTTRT TTRTQ TRTQL RTQLP TQLPP QLPPA LPF
```

```
['LFLPF', 'PLVSS', 'AYXXX', 'XLXPL', 'NNKSC', 'XPLXX', 'TPGDS', 'LHSTQ', 'LPPAY', 'TWFHA', 'SQCVN', 'VYYPD', 'SGWTA', 'QCINL', 'XXVYY', 'XXXXT', 'LDVY
```

```
'LFLPFSXLCXSQYANDALYTYVKVELQHLKDSDLVXXNXXLPXSFVGPHLPDVVFFFXPINVTNXTAFXSSKLIGLKXFPRFVPVSNVAVYFIKTCLHYLLRVIPFXYSFSNMNKMICSVTXAINFDXMLXQXHNLEN
XWMKSPRDMSXCFXNEXKYYVPGKYXLGRGVSFRDXIYVRVFPKLDNLFNIYNLXPHTLLEVIVXYEXSSAISXPSRDNSXPTTLVGFSLSVYXARSYXXVDLAGSLXATFXXXPXXNXXXXFFRIANYXTRXTAGXDF
XXREVFNGTQNVTITXIYSNNNDSYNXYNCTFGYVSAFEXVAYRALKVXYNXPTIXNYCGXFNDLKETTPFRSFGXYXLAQNTINLTXDLDELRDAYTGIXDGIDXDXDQIXVLGXYKSPNYLFXAIADXXERAANNKX
RRLPXVRYXIRDSAPNIYSAYXSPPFPCXXYEGFGCGFSFGLRTGAPAXNVPXLPEXLRFVTLNXLKPRKNKTGTFPVSVFFKGKNGCFADGGAKLNAISGPRGFKQDPSTPIYVLYIXYTDRRVLTWVTILTSSIPFF
EYEAYLLGANCTRNYNETVLYAENLXCQSTTTNVFQQTNQQARXGYTXSHAILANEACCQKXQNGQSECLNVNRMRISIYSRSVASYSTYSMNCSSCKDEVSYTXKXSAXFSYCVTLTTTCELQIKEMDQXDNRGLTEE
TYSARQLNANGELQQVXQKYPQVEAEFNXKAVAKDIXKFIFPFLPGKXCFLVPDRAQLSLAAFINNALPGVIFQLTFVKLIKDXGLGDSIAXAGXANLTAQTILLFLXAXLXGMRNVIQFQVGGAESMQNFXQIAAIAA
NQLLQVQDAEKAXALSQLNSAKKQXLFXAQSDNXIXSLXSAEQXVVXLKXNXTDLNLLXLDQKATFTISLNLQRDTRHNGKGNRLGGRIYTARIL…'
```

Fig 4. 4 New Posssibilities

6. CONCLUSION

On the verge of research, we have built our model using MutaGAN which can convert the Protein Sequence to vectors and it suggests the new mutations possible. As Covid – 19 is being a threat to lives, it is necessary for us to do proper medication. Even though, scientists and researchers working on the drugs, it will be more fortunate to predict the next major variant if it can affect humans as a succeeding wave. In our project, we tried to accomplish the prediction, but we could not get the desired outputs we will work on the algorithms and Mathematics to fight together against the global pandemic as in "Precaution is better than Cure ".

7. REFERENCES:

1. Cilia, E., Teso, S., Ammendola, S. *et al.* Predicting virus mutations through statistical relational learning. *BMC Bioinformatics* **15,** 309 (2014). https://doi.org/10.1186/1471-2105-15-309
2. Ha Young Kim, Dongsup Kim, Prediction of mutation effects using a deep temporal convolutional network, *Bioinformatics*, Volume 36, Issue 7, 1 April 2020, Pages 2047–2052,
3. Salama MA, Hassanien AE, Mostafa A. The prediction of virus mutation using neural networks and rough set techniques. EURASIP J Bioinform Syst Biol. 2016 May 13;2016(1):10. doi: 10.1186/s13637-016-0042-0. PMID: 27257410; PMCID: PMC4867776.
4. Emidio Capriotti, Piero Fariselli, Rita Casadio, A neural-network-based method for predicting protein stability changes upon single point mutations, Bioinformatics, Volume 20, Issue suppl_1, 4 August 2004, Pages i63–i68, https://doi.org/10.1093/bioinformatics/bth928
5. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, Schäffer AA, Brister JR. Virus Variation Resource - improved response to emergent viral outbreaks. National Library of Medicine, SARS-CoV-2 Data Hub , NCBI Virus Nucleic Acids Res. 2017 Jan 4;45(D1):D482-D490. doi: 10.1093/nar/gkw1065. Epub 2016 Nov 28. PMID: 27899678; PMCID: PMC5210549.,
6. Biopython, , Introduction to SeqIO API Documentation, Introduction to SeqIO , Introduction to SeqIO · Biopython
7. Emma Hodcroft, CoVariants,-GISAID data,(28 December,2021), CoVariants
8. Raskin S. Genetics of COVID-19. J Pediatr (Rio J). 2021 Jul-Aug;97(4):378-386. doi: 10.1016/j.jped.2020.09.002. Epub 2020 Oct 7. PMID: 33058776; PMCID: PMC7539923.
9. Wang, R., Chen, J., Gao, K. *et al.* Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. *Commun Biol* **4,** 228 (2021). https://doi.org/10.1038/s42003-021-01754-6
10. Mercatelli Daniele, Giorgi Federico M. Geographic and Genomic Distribution of SARS-CoV-2 Mutations Frontiers in Microbiology ,volume 11, 2020,Pages 1800 ISSN : 1664-302X , DOI: 10.3389/fmicb.2020.01800 , https://www.frontiersin.org/article/10.3389/fmicb.2020.01800

11. KDnuggets , The Ultimate Guide To Different Word Embedding Techniques In NLP,2021 Nov, The Ultimate Guide To Different Word Embedding Techniques In NLP - KDnuggets
12. Machin Library Mastery, What Are Word Embeddings for Text?, What Are Word Embeddings for Text? (machinelearningmastery.com)
13. François Chollet , keras (2015) , GitHub , GitHub repository , commit 5bcac37 , https://github.com/fchollet/keras%7D%7D
14. Analytics Vidhya , Explained Deep Sequence Modeling with RNN and LSTM ,(31 Dec 2020) ,Medium, Explained Deep Sequence Modeling with RNN and LSTM | by Shachi Kaul | Analytics Vidhya | Medium
15. Berman, Daniel & Howser, Craig & Mehoke, Thomas & Evans, Jared. (2020). MutaGAN: A Seq2seq GAN Framework to Predict Mutations of Evolving Protein Populations.
16. Mohamed, T., S. Sayed, A. K. R. A. M. SALAH, and E. H. Houssein, "Long Short-Term Memory Neural Networks for RNA Viruses Mutations Prediction", Mathematical Problems in Engineering, vol. 2021, 2021.
17. Gao K, Wang R, Chen J, Cheng L, Frishcosy J, Huzumi Y, Qiu Y, Schluckbier T, Wei X, Wei GW. Methodology-Centered Review of Molecular Modeling, Simulation, and Prediction of SARS-CoV-2. Chem Rev. 2022 May 20. doi: 10.1021/acs.chemrev.1c00965. Epub ahead of print. PMID: 35594413