Classification of Covid-19 and Covid Variants using DNA Sequence Using Short Time Fourier Transform

Project report submitted in partial fulfillment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

By

Gaurang Khanna (191033)
Anjali Rana (191045)

UNDER THE GUIDANCE OF

Dr. Sunil Datt Sharma

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

May 2023

TABLE OF CONTENTS

DECLARATION

We hereby declare that the work reported in the B.Tech Project Report entitled **"Classification of Covid-19 DNA Sequence using Short Time Fourier Transform"** submitted at **Jaypee University of Information Technology, Waknaghat, India** is an authentic record of our work carried out under the supervision of **Dr.Sunil Datt Sharma.** We have not submitted this work elsewhere for any other degree or diploma.

Gaurang Khanna (191033)
Anjali Rana (191045)

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Sunil Datt Sharma
Date:

Head of the Department/Project Coordinator

ACKNOWLEDGEMENT

Firstly, I express my heartfelt gratitude and gratefulness to almighty God for his divine blessings make it possible to complete the project work successfully. I am really grateful and wish my profound indebtedness to supervisor Dr. Sunil Dutt, Assistant Professor, department of ECE, Jaypee University of Information Technology, Waknaghat. Deep knowledge and keen interest of my supervisor in the field of Machine Learning to carry out this project . His endless patience , scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, reading many inferior drafts and correcting them at all stages have made it possible to complete this project successfully.

I would also generously welcome each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patience of my parents.

Gaurang Khanna
191033
Anjali Rana
191045

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Coronavirus(Covid-19) is a severe pandemic that has affected millions around the world. It was an unforeseen viral infection that happened in the year 2020 in the Wuhan city of China. The fast spread of this danger affected the health of people across the globe. In the previous twenty years, many communicated diseases are marked like SARS-CoV, severe acute respiratory syndrome coronavirus in 2002-2003 and recently middle east respiratory syndrome coronavirus (MERS-CoV) which first existed in 2012. With the spread of Covid-19, various mutations (when the basic structure of any species changes itself) were introduced, these mutations were able to infect a person with serious illness. Some of the variants were highly transferable making the variants more infectious. In this study we have tried to differentiate between MERS-COV and SARS-CoV and we have used the dataset of strain XBB.1.5, BA.5.2.1 with the use of genome datasets from NCBI for making detection procedures smooth and increasing the accuracy of virus detection in quick time. The selected datasets contain 20 genomic sequences for Sars and mers each whereas these variants contain 15 datasets of each respectively. Then the classification between sars-cov - mers-cov and between the strains XBB.1.5, BA.5.2.1 is being performed using a short fourier transform in MATLAB. The results obtained are in the form of threshold values from which we can differentiate between the sars- mers and between the strains of covid-19(XBB.1.5, BA.5.2.1).

CHAPTER 1

1.1 INTRODUCTION

Coronavirus are RNA viruses (single positive stranded) which cause infection to mammals. The major proteins involved are Silk Protein, Nucleocapsid, Membrane and Envelope. The spike protein present gives these viruses a crown-like look when seen through an electron microscope[15,16]. This structure is called the colonum.

The ICTV considered these coronaviruses are Nidovirals members which comes under the category of Cornidovirineae. Basically coronavirus falls under two subcategories called Letovirinae and Orthocoronavirinae[12]. It is further divided into four types: gammaCoV, alphaCoV, betaCov and deltaCov. The Alpha and Beta strain are able to hit mammals while the other two lead to avian infection. Whereas mers-CoV and sars-CoV inflates extreme lung infections.

The year 2019 saw a different variant of covid which was carried in Wuhan that transmitted from animals to humans as well. The World Health Organisation (WHO) named it as Coronavirus Disease 2019 or popularly known as Covid-19. The ICTV committee named this deadly virus as SARS-COV-2.[13,14] Coronavirus is a highly infectious disease, it has a fatal effect on the world resulting in a high number of deaths worldwide. Covid is a contagious virus that spreads rapidly in a very short time of span. In the start of the pandemic, there were comparatively low numbers of infections resulting in less chances for mutations. As time passed, the evolution of covid -19 genome led to various mutations. The threat of covid began when this virus started mutating itself into various strains, many of these strains were dangerous, highly infectious and transmitted at a very quick and frightening rate. As of May 03, 2023, a total of 765,222,932 cases of the COVID-19 have been registered worldwide, which also include a large number of deaths, i.e. 6,921,614. Since the widespread of the virus, many variants have been discovered and a few of them are a matter of great concern such as Alpha(B.1.17), Beta(B.1.351), Gamma(P.1), Omicron(B.1.1.529). Covid

variants can differ in their characteristics. For example, some may spread quickly, some may show no effect or may be resistant to available treatments. So, WHO has classified the variants as variants of interest, variants of concern and variants under monitoring.

1.1.1 Variants

We have done the classification of two variants for our project and they are named **XBB.1.5** and **BA.5.2.1**. Both of these variants are the sub-variants of the omicron covid variant.

- **XBB.1.5-** Since the omicron variant is dominating other variants, it has mutated into multiple variants. One of the sub variants is XBB.1.5, scientists also called it “Kraken”. This is a subdivision of XBB, a combination of two BA.2 variants, with an addition of a spike receptor. This variant is believed to bind cells and become more transmissible. This spread so quickly, it was reported XBB.1.5 accounted for 40% cases of covid in the US, making it the most dominant strain. Its symptoms are quite similar to other variants like fever, body ache, chills, fatigue, sore throat, running nose, nausea or vomiting.

- **BA.5.2.1-** Another subvariant of omicron is BA.5.2.1 and the first case was found in the city of Shanghai. This virus accounted for approximately 53% covid cases. This has a better ability to dodge the immune system. The common symptoms are cough, nasal congestion. With some people it was claimed that the strain is asymptomatic, making it hard to detect the variant.

With the evolution of Sars - covid, various variants have been discovered that differ in genomic sequence from the virus found in the year 2019 in Wuhan. These variations differ in terms of rate of transmission, clinical symptoms and vaccine efficiency.

1.2 Objective

The Covid19 pandemic is such a huge damage to our society and is still one of the main concerned issues around the globe. Infection and mortality rates are intermediate compared to other deadly viruses. As a result, the severity of the disease causing coronavirus was greatly underestimated by society at the beginning of the outbreak[14]. To diagnose this deadly virus, many methods are introduced like molecular diagnosis, PCR based testing methods, isothermal nucleic acid amplification based method and many more. It was not very difficult to detect covid affected patients, but problem arises when the strains were introduced, it was a challenging task to detect the infected patient suffering from which type of strain so that the treatment can be done accordingly. These newly discovered strains were posing hurdles to the health care system as these strains were difficult to detect and classify. The variants of concern are more contagious and causing deaths worldwide. The most recent variant is XBB.1.5. which is considered to be dangerous. So, it was necessary to classify the covid and identify covid-19 variants without human interaction. It can be done with computer methods and with the help of artificial intelligence systems.

In this study, for the first half we have differentiated between covid and non covid, in the second half we have differentiated between variants of covid-19, i.e. XBB.1.5 and BA.5.2.1, and these two variants are the most recent variants discovered by the World Health Organisation (WHO). The proposed system is to differentiate the coronavirus and its variants with minimal human interaction and to differentiate between the covid and non covid and between covid variants. This system is more reliable and depends on GSP and the classification algorithm.

1.3 Motivation

Covid-19 pandemic has left such a huge impact on the globe in almost every sector of the society. It has tremendously affected both healthcare and economic sectors worldwide. This pandemic was so devastating that it averaged around 6000 deaths per day internationally. The prognosis prediction for the virus infected patients was very difficult at first. As of May 06, 2023, there are still 20,668,199 active cases globally and 30,041 in India, including the emergence of new strains of coronavirus. As these variants are sub-lineage of SARS covid, they have very similar characteristics which make them difficult to detect and classify among them. The variants do not show much disparity among themselves, which is still a troublesome work for the scientists to find the cure of that particular variant. The most common and fast RT-PCR test is able to detect only one SARS-Cov- omicron variant (B.1.1.529), which is also having an accuracy of 70-80% for finding the variant. Moreover, there have not been many researches as well till the date for the same concerned issue. So, we aim to differentiate between various covid and non covid, and different strains of covid-19 (XBB.1.1, BA.5.2.1) using a short time fourier transform in MATLAB.

1.4 Technical Requirements

The technical requirements for this project are -

1. MATLAB

CHAPTER 2

2.1 Literature Review

The outbreak of Coronavirus Disease has caused a serious threat to public health worldwide. This pandemic is a great source of research for many big data analytics and artificial engineers. Ample of research has been done to diagnose this deadly virus. In 2020, Barstugab et al.[1] researched this classification method using various types of techniques like Computed Tomography images(CT), feature extraction method and SVM. Basu et al [2] found some characteristics like respiratory X-Ray scans and took DETL for the detection of covid-19. Ozturk et al [3] used raw chest X-Ray images and dark covid net model as a classification measure to make two different types of classifications - binary classification and multi-class classification. Elasnaoui et al [4] used both X-Ray and CT images to find out bacterial pneumonia, coronavirus, covid 19 and used various deep learning models. The accuracy rates for the above methods are as follows - 99.6%, 95.3%, 98.08% & 92.18 respectively.

A dataset named covid-x and deep CNN learning model, covid net was introduced by Wang et al [5] to diagnose the patients from chest X-Ray images. These X-Ray images were collected from the perspective of covid positive, normal and pneumonia viral. This proposed system is a combined and modified form of the available dataset. This system achieved an overall accuracy of 92.4%. Khan et al [6] proposed a CNN based covid 19 detection model by X-Ray and CT scan images from cohen jp and chest X-Ray images pneumonia dataset. This experiment was conducted by binary 3 class and 4 class classification. The detection model was the exception model and was pre-trained on image net dataset. The results show accuracy of binary , 3 class and 4 class were 99%, 95% & 89.6% respectively. Maghdid [7] also proposed a CNN based deep learning for covid 19 detection using X-Ray and CT images and they were able to achieve accuracy of 98%.

Ali Narin[8] has developed an automatic detection system as a diagnosis of covid-19. In this research, three different CNN based models (resnet 50, inception V3 and inception - resnet V2) have been proposed to detect this deadly virus using chest X-Ray radiograph. In this paper, the classification performance accuracy between three CNN models is being discussed.

In [9] the authors proposed a three indices based model to predict the mortality risk. They developed a prognostic prediction model based on the XG boost machine learning algorithm to predict the mortality risk in the patient. They found a clinical root which is simple to check and assess the risk of death. While another article [10] presented a comparative analysis of machine learning models to predict the outbreaks of covid 19. This article was completely based on the outbreak of cases in various countries.

On the other hand, [11]the diagnosis were done on some analysis carried out in medical centers like RT-PCR test and CT scan. These techniques are accurate to an extent but they still have many setbacks. There are inadequate test kits and difficulty of the test themselves such as requirement for suction devices as the number of patients were many, and due to expensive expenditure and long period of time to get results.

[23]Another study shows the development of new PCR based tests for differentiating and classifying among variants. These tests are considered to be more efficient and easier. [24] Some developed a set of four multiplex mutation PCR based tests that can detect VOC. [25]Some of the researchers created a multiplex reverse transcriptase quantitative PCR (RT-qPCR) assay that can correctly identify and can distinguish among the newly evolving B.1.1.7 and B.1.351 SC-2 variants. In this study, they combined 4 reactions—one of them find SC-2 RNA individualistically of the strain, one which is able to detects the D3L mutation, which is only specified to a particular variant named as B.1.1.7, third one that detects the 242 to 244 deletion, which is dedicated to the variant B.1.351, and the last one, which identifies the human RNAseP gene, serving as an autogenous control for RNA extraction. [26]This study about variants, selected three particular genomic regions of the following virus *N*, Orf1ab and *S*. The sequencing which is used to confirm

positive PCR output is Whole-genome next-generation sequencing. After that, positive outputs were mapped using the postal samples which enabled the real-time monitoring of the spread of a new virus throughout the UK. [27]This describes a simple, fast, and highly efficient reverse transcriptase PCR (RT-PCR) melting-temperature (*Tm*) screening research that identifies the initial three crucial VOCs. The RT-PCR is based on sloppy molecular beacon (SMB) in which the study was designed to rectify and detect the SARS-CoV-2 N501Y (A23063T) and E484K (G23012A) mutations and their corresponding sublinages. After the RT-PCR test, the major VOCs were identified by a characteristic *Tm* of each SMB. The enhanced result and the testing was performed with RNA from SARS-CoV-2 USA WA1/2020 (wild type [WT]), B.1.1.7, and B.1.351 variant strains.[28]This following system verified and implemented an N501Y-specific PCR to rapid screening for some of the VoCs, which are then proved using an amplicon sequencing or whole genome sequencing (WGS). A total of 13,387 VoCs have been discovered since the detection of the first Swiss case, with 4194 being B.1.1.7, 172 B.1.351, and 7 P.1. and explained the nationwide coordination and execution process among some labs and public health institutions, the first results of the N501Y-specific variant screening, and the phylogenetic analysis of all available WGS data, that collaboratively identified the initial introductory events and successive community transmitting of the VoCs.

2.2 Problem Definition

In this research, we aimed to classify between SARS and MERS using Short Term Fourier Transform (STFT) as a feature extraction method in MATLAB in the first half. For the second half, we aim to classify between the two strains of coronavirus.

2.3 Solution

2.3.1 STFT

The Short Time Fourier Transform divides the time-domain input signal into multiple overlapping blocks with the help of a window function multiplied by the signal itself. Sliding windows or the window functions are mathematical functions which have zero values outside of the selected range. The concluding STFT shows the spectral content of the signal at each correlating time period as each block is engaged with different time periods. The spectral content of the signal is obtained over different time intervals on moving the sliding window. This indicates that the STFT is basically a function of both the time and the frequency domains. The magnitudes of the STFT coefficients give the magnitude time-frequency spectrum, whereas the phases of STFT coefficients give the phase time-frequency spectrum. STFT is basically used to find out the frequency of a nonstationary signal which changes over time. We have selected this method because it has smaller time frames resulting in more accurate output.

Mathematical definition of STFT

$X□(\omega) = = \sum_{n=-\infty}^{\infty}$ ▒ $[\![x(n)w(n-mR)e^{-j\omega n}]\!]$

2.3.1 Materials and Methods

The below block diagram gives a view of the research. Firstly, the selected DNA sequence is taken from the National Centre for Bioinformatics and then this database is being converted to numeric values. The converted database is now classified using a short fourier transform.

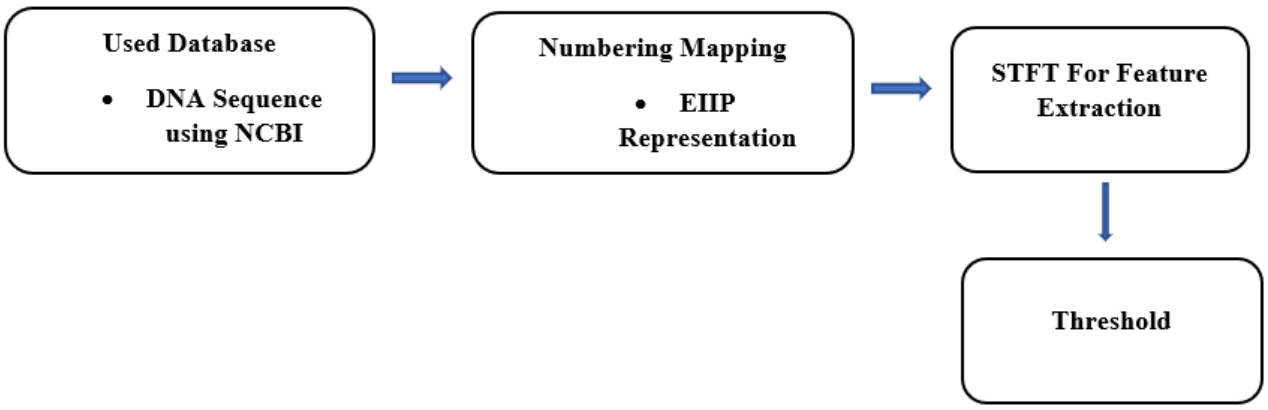


Fig 1. Flow Chart of the Methodology

2.3.2 Database

The National Centre for Biotechnology Information (NCBI) at the National Institutes of Health is the main source of nucleotides dataset for many diseases which are related to genetic.The dataset used is also known as Genbanks. In this study we have used 20 genomes for epidemic SARS-CoV and MERS each, and each of them has 29,000 nucleotides and for Covid variants XBB.1.5 and BA.5.2.1 we have used 15 genomes each of them has 29,000 nucleotides. RNA is taken out for the entire genomic sequence of the infectional disconnect[17,18]. Initially, RNA was taken from a clear cellular culture resilient and randomly opened up cDNA arranged by (SISPA). Sequencing was performed by combining ONTI short-read sequencing. Genomic gathering of the betaCoV/australia/vic/01/2020 genome was agreed and reference-guided strategies.

2.3.3 Numbering Methods (mapping)

The electron-ion interaction pseudopotential(EIIP) representation is one of the mapping methods we have used, as this mapping method as the values are published to provide optimal mapping for spectral analysis of DNA sequences. We know that DNA is made up of four basic blocks known as the nucleotides which are Adenine(A), Thymine(T), Guanine(G), Cytosine(C). This function replaces each nucleotide sequence with electron-ion interaction. There are numerous methods to convert the nucleotides sequences string 'A' = 0.1260, 'G' = 0.0806, 'C' = 0.1340 and 'T' = 0.1335 as numeric values[19-22].

2.3.4 Feature extraction using STFT

Feature extraction is basically dimensionality reduction process in which a large number of data is reduced into smaller groups, making the process more simple. We have used the Short time fourier transform(STFT) as a feature reduction method for our project. Short-time Fourier transform (STFT) is a mathematical tool, which is used to process the non-stationary signals. It maps the information present in the signal in time and frequency domain simultaneously. In this work, STFT has been used to extract the feature i.e concentration measure of the DNA sequences of the covid-19.

2.3.5 Thresholding

Appropriate threshold has been selected experimentaly to categorize the DNA sequences of the COVID-19 and MERS-COVID. In this study, we have classified between the SARS-CoV and

MERS using a short fourier transform. The Short-time Fourier transform (STFT) is used to examine how the frequency content of transient signals changes over time. The STFT magnitude squared is known as the spectrogram time-frequency representation of the signal. Since the frequency resolution of STFT is the same for all positions in the spectrogram, STFT offers only a sub-optimal compromise between time resolution and frequency resolution.

Results

SARS-Cov and MERS Classification

In the SARS-CoV and MERS classification process we succeeded to recognize all the databases using fourier transform and obtained the concentration values for both the datasets. Also we get the signal energy graph for both the variants.

Example 1

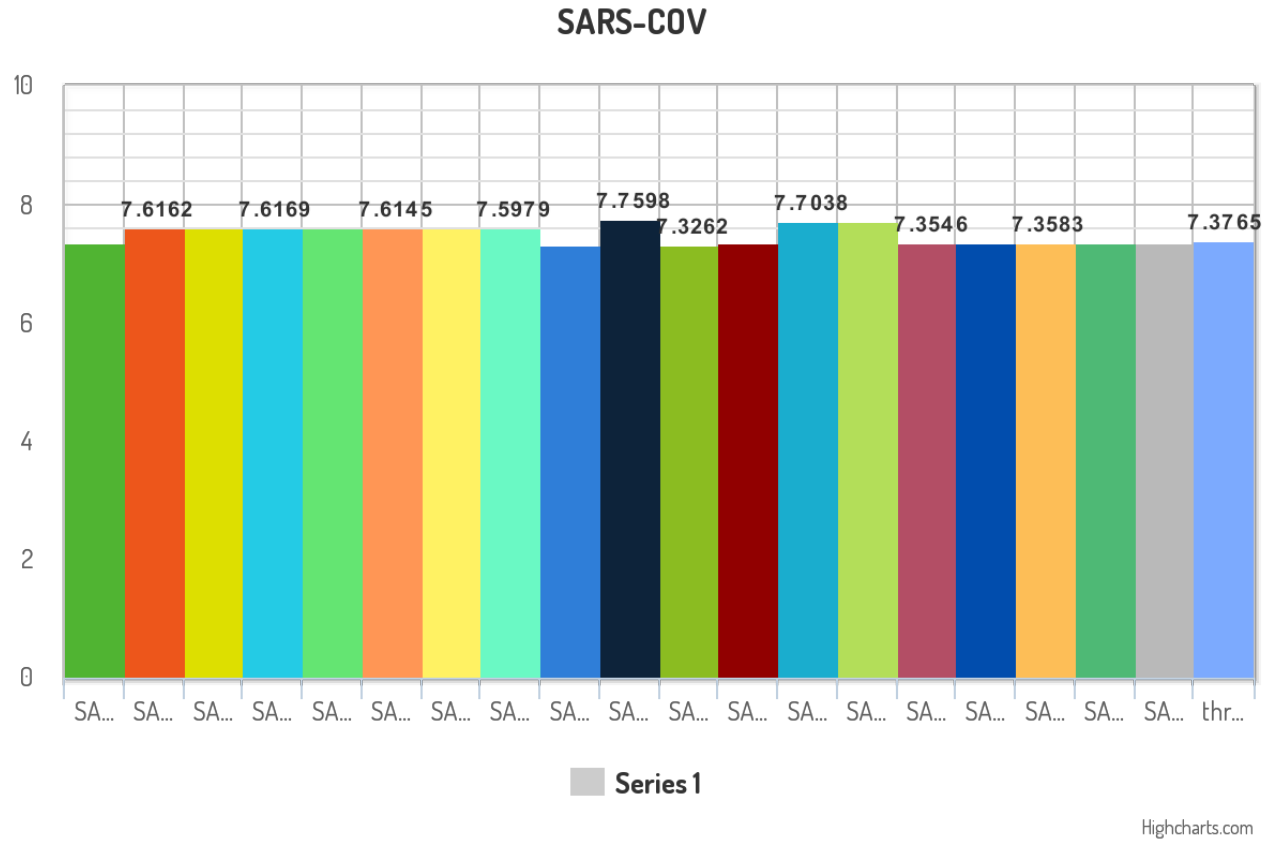


Fig 2 Measure Of Signal Energy Graph for SARS-CoV

Example 2

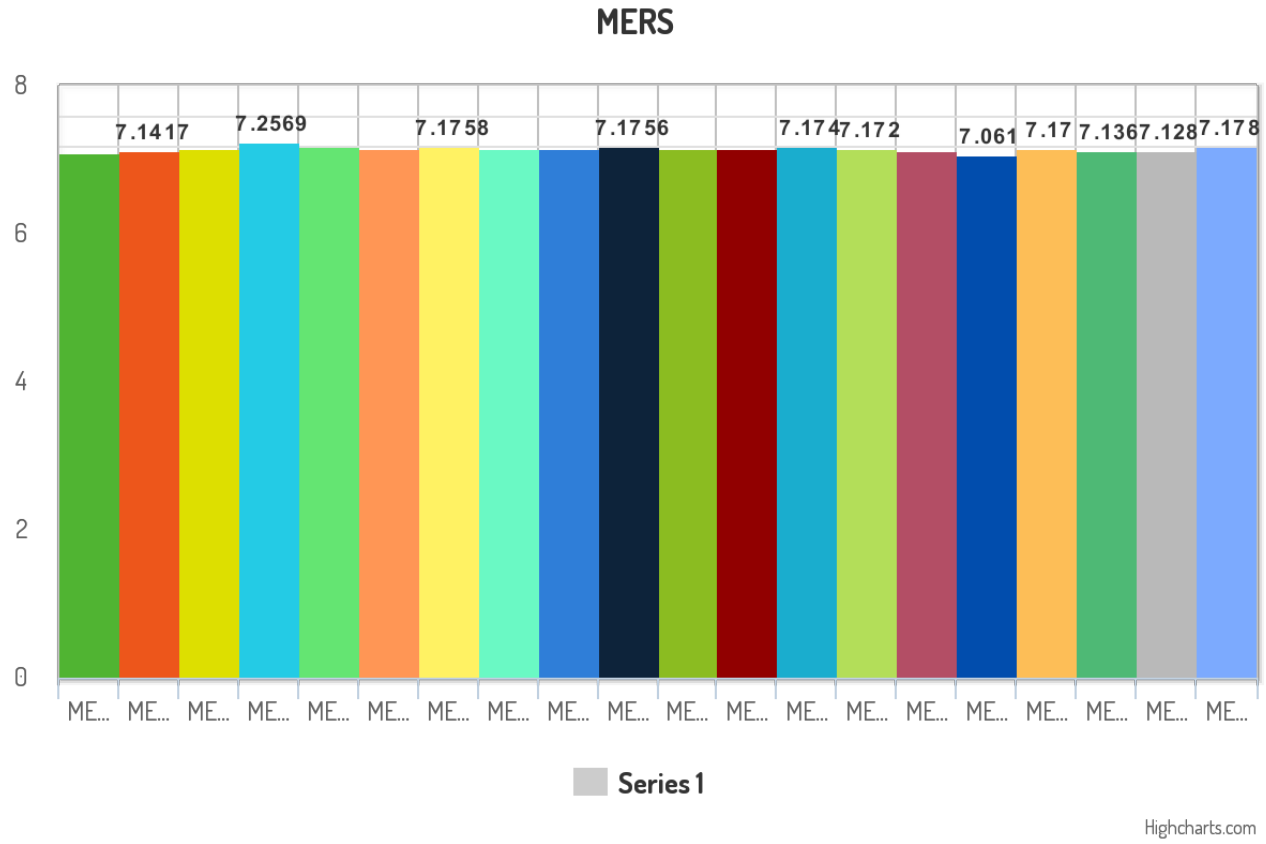


Fig 3 Measure of Signal Energy Graph for MERS

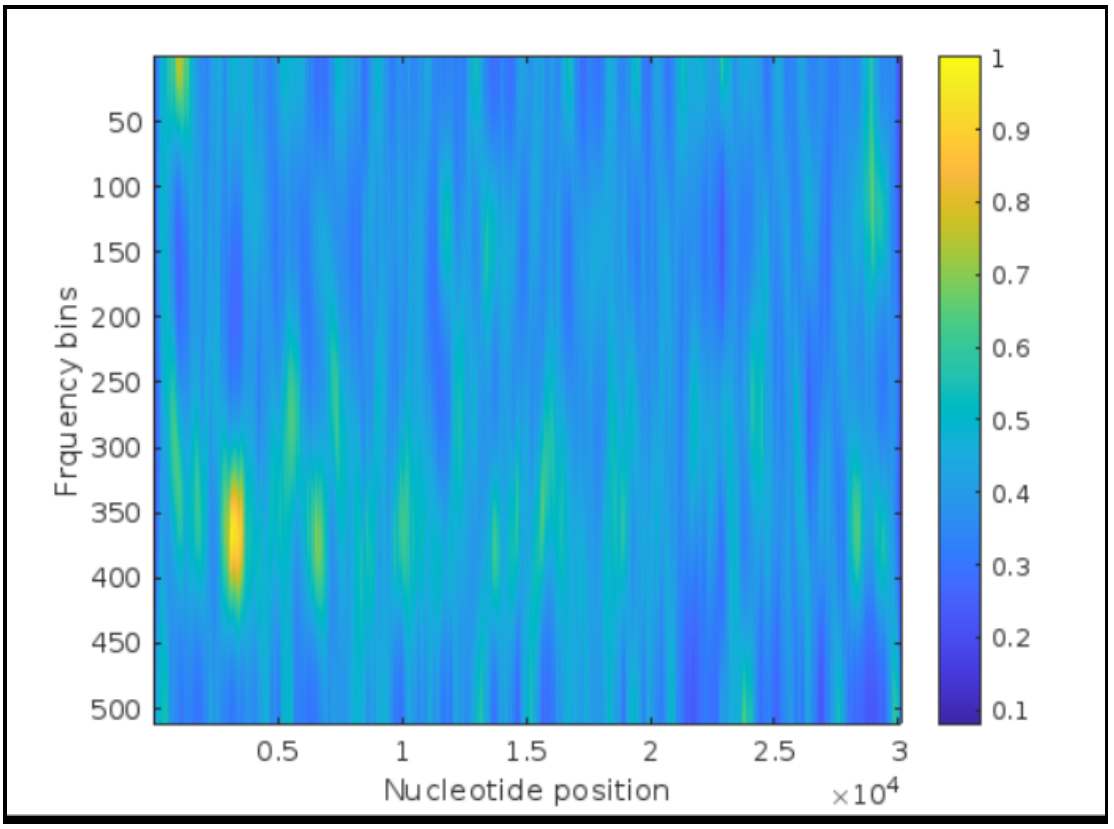


Fig 4 Spectrogram for MERS DNA Sequence

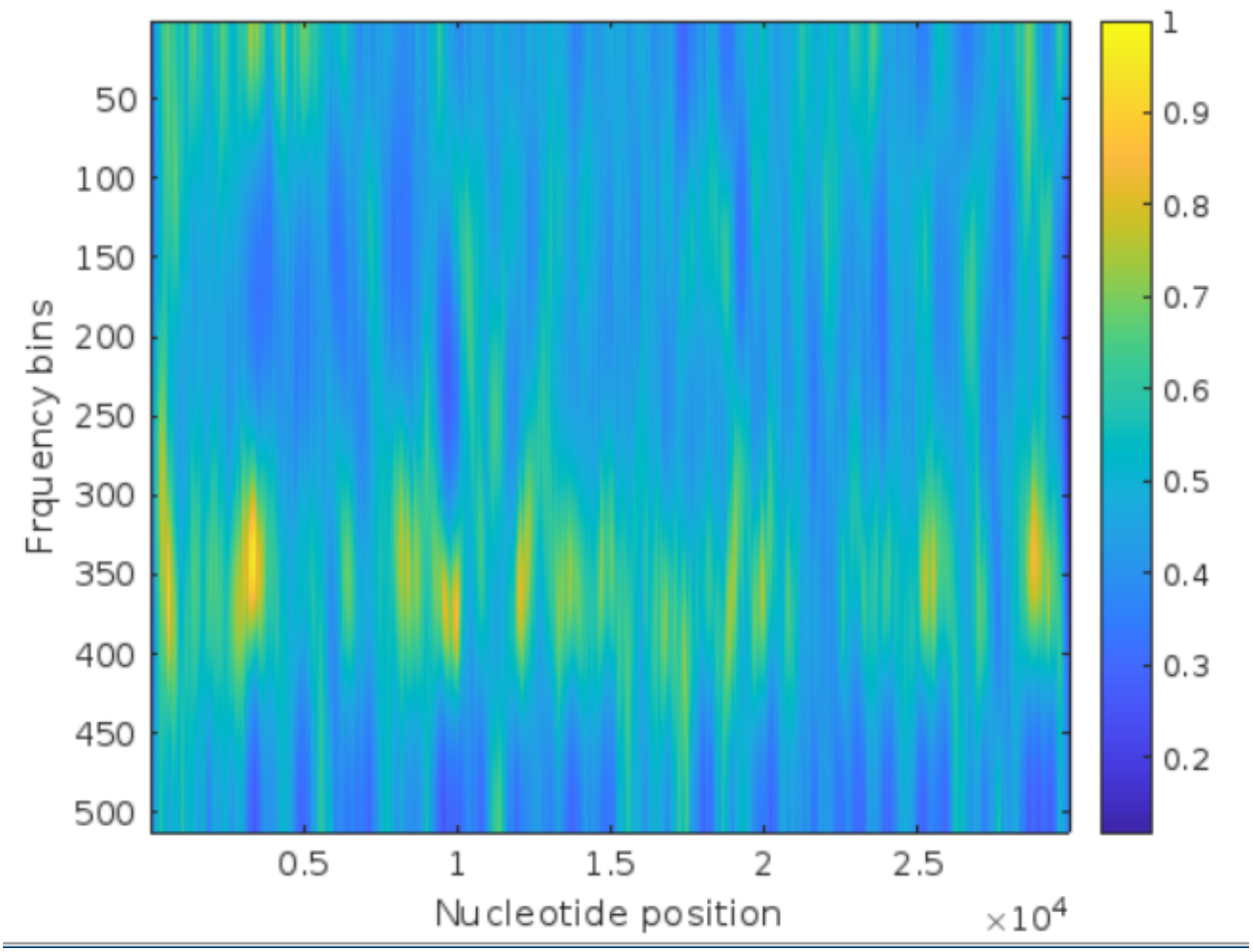


Fig 5 Spectrogram for SARS DNA Sequence

Table 1

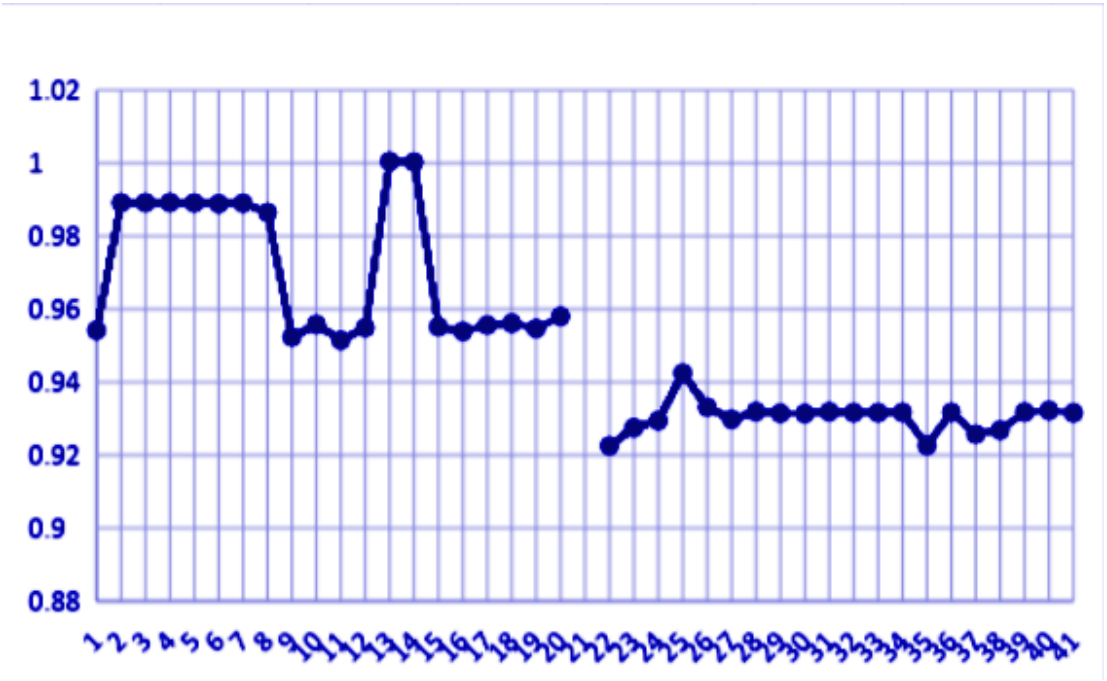
S.No	Accesion No.	DNA Sequence	Concentration Value
1	MZ722015.1	SARA-CoV	7.347*10^-8
2	MZ472096.1	SARA-CoV	7.6162*10^-8
3	MZ472097.1	SARA-CoV	7.6165*10^-8
4	MZ472098.1	SARA-CoV	7.6169*10^-8
5	MZ472099.1	SARA-CoV	7.6157*10^-8
6	MZ472103.1	SARA-CoV	7.614*10^-8
7	MZ472104.1	SARA-CoV	7.615*10^-8
8	MZ472095.1	SARA-CoV	7.5957*10^-8
9	MZ353007.1	SARA-CoV	7.3328*10^-8
10	MZ472102.1	SARA-CoV	7.3598*10^-8
11	MZ353013.1	SARA-CoV	7.3262*10^-8
12	MZ022251.1	SARA-CoV	7.3525*10^-8
13	MZ472100.1	SARA-CoV	7.7038*10^-8
14	MZ472101.1	SARA-CoV	7.7020*10^-8
15	MZ185542.1	SARA-CoV	7.3546*10^-8
16	MZ283890.1	SARA-CoV	7.3450*10^-8
17	MZ184982.1	SARA-CoV	7.3583*10^-8
18	MZ368207.1	SARA-CoV	7.3622*10^-8
19	OL369765.1	SARA-CoV	7.3510*10^-8
20	MZ034468.1	SARA-CoV	7.3765*10^-8

Table 1. Concentration values of SARS-CoV-19

Table 2

S.No	Accesion No.	DNA Sequence	Concentration Value
1	MG987420.1	MERS-CoV	7.1035*10^-8
2	MG987421.1	MERS-CoV	7.1417*10^-8
3	ON325306.1	MERS-CoV	7.1567*10^-8
4	MK129253.1	MERS-CoV	7.2569*10^-8
5	MT361640.1	MERS-CoV	7.1851*10^-8
6	OP654179.1	MERS-CoV	7.1595*10^-8
7	KF192507.1	MERS-CoV	7.1758*10^-8
8	KJ813439.1	MERS-CoV	7.1722*10^-8
9	KJ829365.1	MERS-CoV	7.1716*10^-8
10	KP209306.1	MERS-CoV	7.1756*10^-8
11	KP209307.1	MERS-CoV	7.1735*10^-8
12	KP209308.1	MERS-CoV	7.1734*10^-8
13	KP209309.1	MERS-CoV	7.1740*10^-8
14	KP209310.1	MERS-CoV	7.1042*10^-8
15	KP209311.1	MERS-CoV	7.174*10^-8
16	KP209312.1	MERS-CoV	7.1280*10^-8
17	KP209313.1	MERS-CoV	7.1365*10^-8
18	KP223131.1	MERS-CoV	7.1748*10^-8
19	KT026454.1	MERS-CoV	7.1780*10^-8
20	KT156560.1	MERS-CoV	7.1730*10^-8

Table 2. Concentration values of MERS-CoV



Graph1. classification of covid and non covid

Variants Classification

For the classification of the two different omicron covid variants, we succeeded to recognise all the datasets for each variant using short time fourier transform and obtained the concentration values for each. Also we have the signal energy graph for both of them.

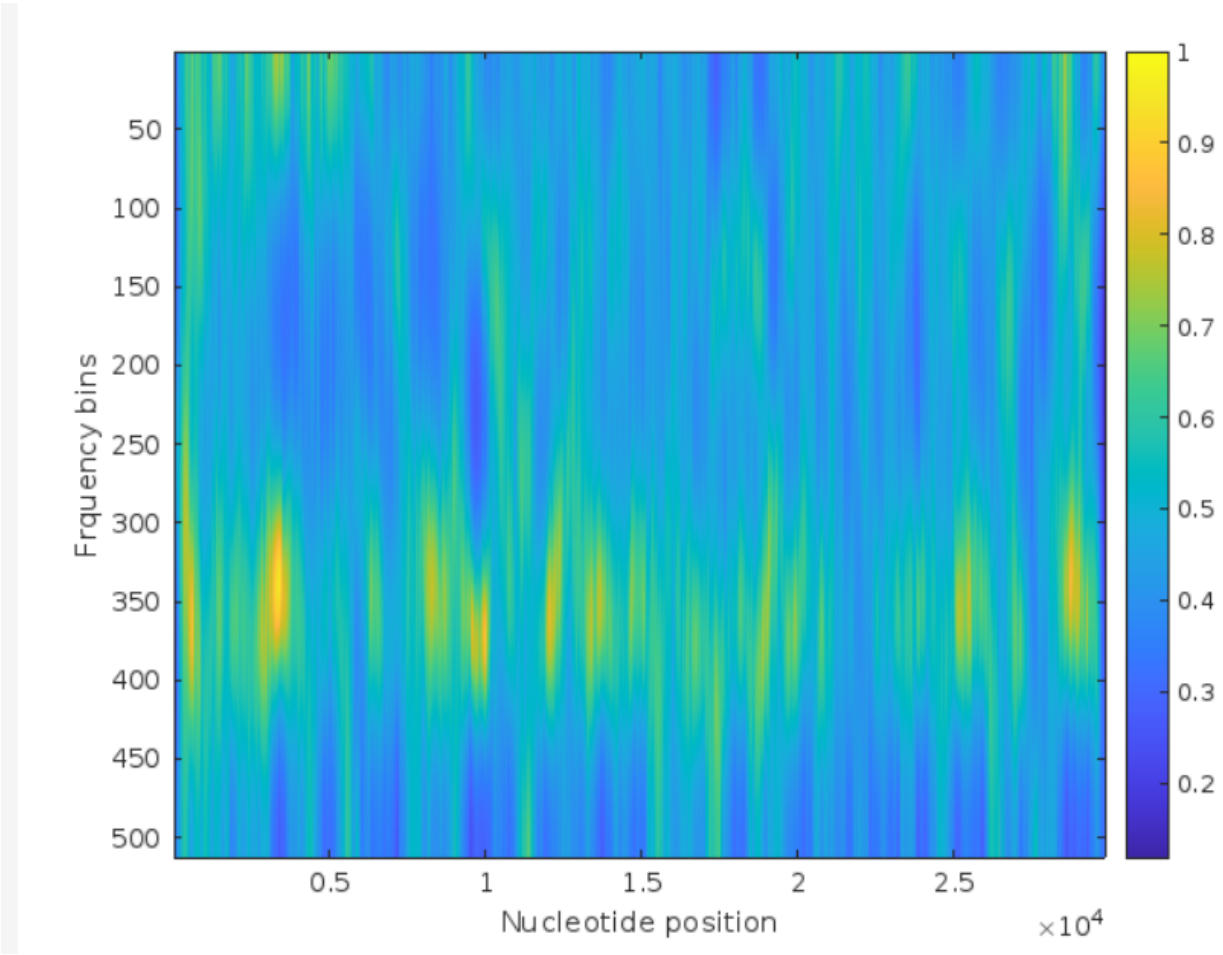


Fig 6. Spectrogram for XBB.1.5 DNA Sequence

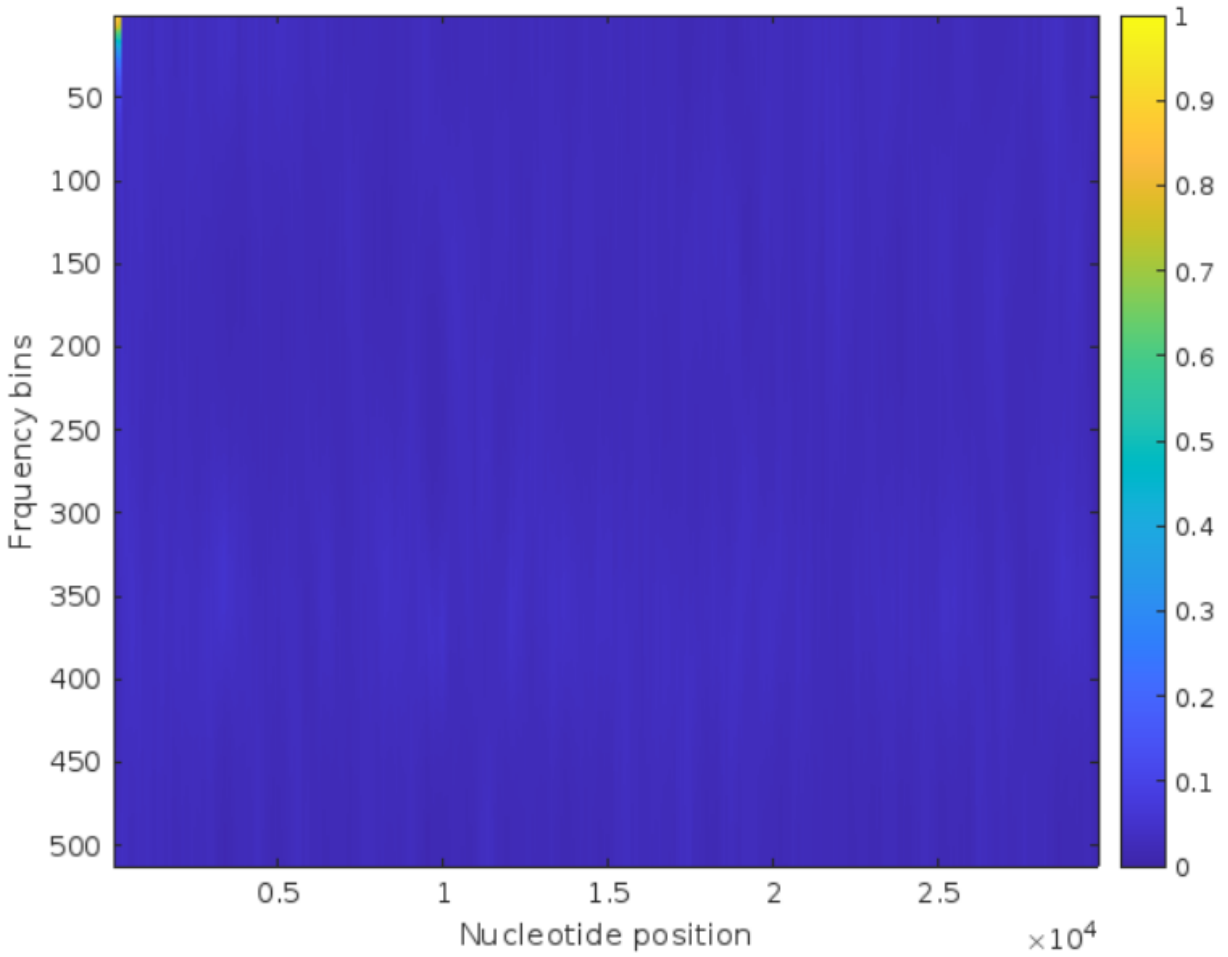
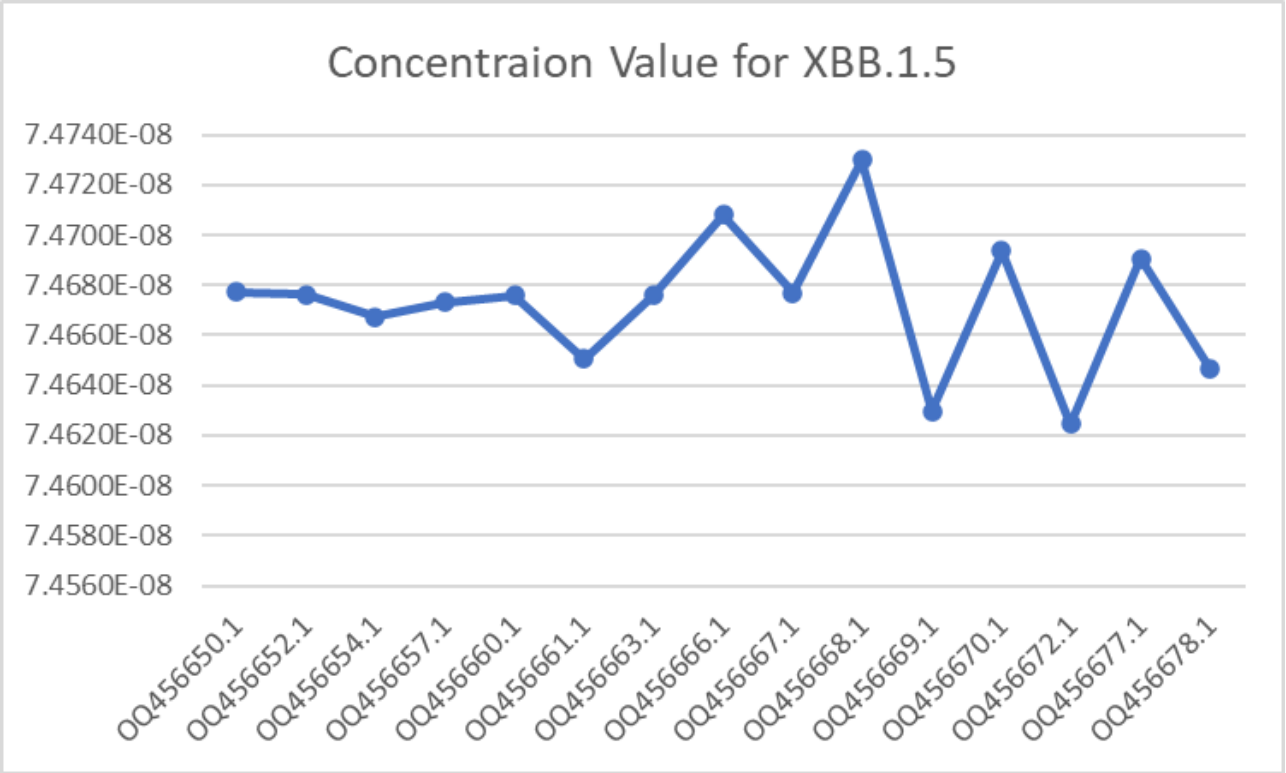


Fig 7. Spectrogram for BA.5.2.1 DNA Sequence

Table 3

S.No.	Accession No.	Variant	Concentration Value
1	OQ456650.1	XBB.1.5	7.4677E-08
2	OQ456652.1	XBB.1.5	7.4676E-08
3	OQ456654.1	XBB.1.5	7.4667E-08
4	OQ456657.1	XBB.1.5	7.4673E-08
5	OQ456660.1	XBB.1.5	7.4676E-08
6	OQ456661.1	XBB.1.5	7.4651E-08
7	OQ456663.1	XBB.1.5	7.4676E-08
8	OQ456666.1	XBB.1.5	7.4708E-08
9	OQ456667.1	XBB.1.5	7.4677E-08
10	OQ456668.1	XBB.1.5	7.4730E-08
11	OQ456669.1	XBB.1.5	7.4630E-08
12	OQ456670.1	XBB.1.5	7.4694E-08
13	OQ456672.1	XBB.1.5	7.4625E-08
14	OQ456677.1	XBB.1.5	7.4691E-08
15	OQ456678.1	XBB.1.5	7.4647E-08

Table 3 . Concentration values for Variant XBB.1.5

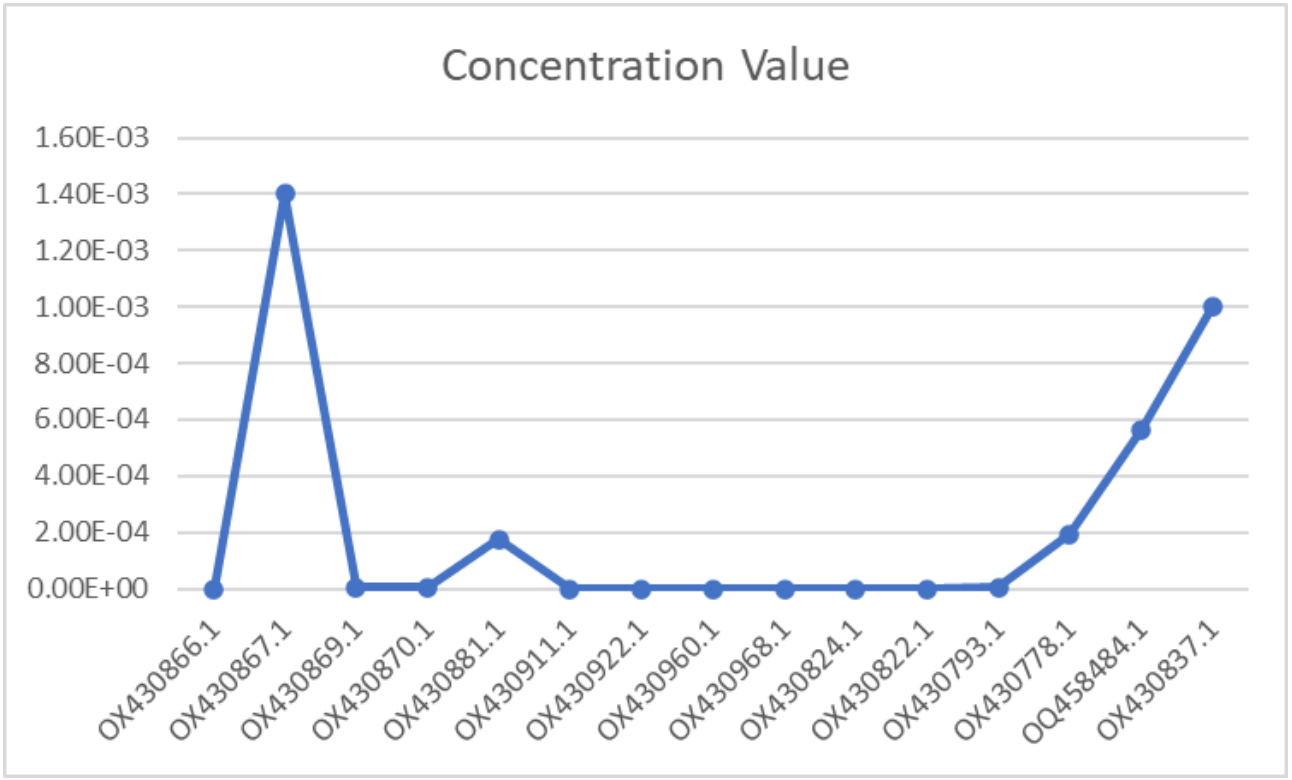


Graph 2. Representation of Concentration values of XBB.1.5

Table 4

S.No.	Accession No.	Variant	Concentration Value
1	OX430866.1	BA.5.2.1	1.03E-07
2	OX430867.1	BA.5.2.1	1.40E-03
3	OX430869.1	BA.5.2.1	5.00E-06
4	OX430870.1	BA.5.2.1	4.85E-06
5	OX430881.1	BA.5.2.1	1.75E-04
6	OX430911.1	BA.5.2.1	1.86E-07
7	OX430922.1	BA.5.2.1	1.32E-07
8	OX430960.1	BA.5.2.1	1.11E-07
9	OX430968.1	BA.5.2.1	9.96E-08
10	OX430824.1	BA.5.2.1	NaN
11	OX430822.1	BA.5.2.1	9.06E-08
12	OX430793.1	BA.5.2.1	5.29E-06
13	OX430778.1	BA.5.2.1	1.92E-04
14	OQ458484.1	BA.5.2.1	5.65E-04
15	OX430837.1	BA.5.2.1	1.00E-03

Table 4. Concentration values for variant BA.5.2.1



Graph 3.Representation of concentration values of BA.5.2.1

Conclusion

After the spread of the COVID-19 virus, many researchers around the world began to study and analyze medically to get medicine for this terrible disease as soon as possible. Numerous tests, including RT-PCR and RRT-PCR, can diagnose COVID-19 instances, however they have several limitations: Lack of test kits, the requirement for suction equipment, the expensive cost, the lengthier time it takes to receive results, and patient pain. Genetic illnesses using genomic signals include coronaviruses. The fundamental strategies needed for the aforementioned investigation are classifier selection and processing methods. In this study, an automated technique for identifying COVID-19 and its variants, differentiating it from other coronaviruses including SARS-CoV , MERS and XBB.1.5, BA.5.2.1 is presented. This technique facilitates quick COVID-19 diagnosis. The process will proceed with the DNA centrifuge to isolate the DNA sequence. It only takes a short while. All outcomes are respectable and satisfying. For the best and most effective diagnosis method, this is best suited for classification systems. To accomplish new objectives for Covid-19, alternative functions or classifiers will be chosen in future work.

REFERENCES

[1] Barstugan M, Ozkaya U, Ozturk S "Coronavirus(Covid-19) classification using CT images by Machine learning methods" IEEE Trans. Biomed. Eng arXiv: vol -203 , pp 09424, oct, 2020 preprint.

[2] Basu S, Mitra S, Saha N " Deep learning for screening of Covid-19 using chest X-ray images" Appl. Phys. Lett., vol. 91, pp 10507, Nov, 2020.

[3] Ozuturk T, Talo M "Automated detection of covid-19 cases using deep neural networks with X-ray images" comput , vol.45,pp sep,2021.

[4] A. Le, A. Markopoulou, and M. Faloutsos, "PhishDef: URL names say it all," in Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM), Sep. 2010, pp. 191–195.

[5] Lu, Guangwen, Qihui Wang, and George F. Gao. "Bat-to-human: spike features determining 'host jump'of coronaviruses SARS-CoV, MERS-CoV, and beyond." *Trends in microbiology* 23.8 (2015): 468-478.

[6] Khan AI, Shah JL, Bhat MM. CoroNet: A deep neural network for detection and diagnosis of COVID-19 from chest x-ray images. Comput Methods Programs Biomed. 2020 Nov;196:105581. doi: 10.1016/j.cmpb.2020.105581. Epub 2020 Jun 5. PMID: 32534344; PMCID: PMC7274128.

[7] Maghdid, Halgurd & Asaad, Aras & Ghafoor, Kayhan & Sadiq, Ali & Khan, Khurram. (2020). Diagnosing COVID-19 Pneumonia from X-Ray and CT Images using Deep Learning and Transfer Learning Algorithms.

[8] Narin, A., Kaya, C. & Pamuk, Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. *Pattern Anal Applic* 24, 1207–1220 (2021).

[9] Ma X, Ng M, Xu S, Xu Z, Qiu H, Liu Y, Lyu J, You J, Zhao P, Wang S, Tang Y, Cui H, Yu C, Wang F, Shao F, Sun P, Tang Z. Development and validation of prognosis model of mortality risk in patients with COVID-19. Epidemiol Infect. 2020 Aug 4;148:e168. doi: 10.1017/S0950268820001727. PMID: 32746957; PMCID: PMC7426607.

[10] Abirami, R.S., Kumar, G.S. Comparative Study Based on Analysis of Coronavirus Disease (COVID-19) Detection and Prediction Using Machine Learning Models. *SN COMPUT. SCI.* 3, 79 (2022).

[11] Udugama, B., et al. "Diagnosing COVID-19: the disease and tools for detection. ACS Nano 14: 3822–3835." *Link: https://bit. ly/2VbmlOD* (2020).

[12] Pradhan, Prashant, et al. "Uncanny similarity of unique inserts in the 2019-nCoV spike protein to HIV-1 gp120 and Gag." *BioRxiv* (2020).

[13] Guo, Yan-Rong, et al. "The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak–an update on the status." *Military medical research* 7.1 (2020): 1-10.

[14] Pan, Lei, et al. "Clinical characteristics of COVID-19 patients with digestive symptoms in Hubei, China: a descriptive, cross-sectional, multicenter study." *The American journal of gastroenterology* 115 (2020).

[15] Cascella, M., et al. "StatPearls [Internet] StatPearls Publishing; Treasure Island (FL): 2020." *Features, evaluation and treatment coronavirus (COVID-19)[updated 2020 Apr 6] https://www. ncbi. nlm. nih. gov/books/NBK554776/Available from:[Google Scholar]*.

[16] Zhou, Peng, et al. "A pneumonia outbreak associated with a new coronavirus of probable bat origin." *nature* 579.7798 (2020): 270-273.

[17] Lewandowski, Kuiama, et al. "Metagenomic nanopore sequencing of influenza virus direct from clinical respiratory samples." *Journal of clinical microbiology* 58.1 (2019): e00963-19.

[18] Kafetzopoulou, Liana E., et al. "Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples." *Eurosurveillance* 23.50 (2018): 1800228.

[19] Das, J., and S. Barman. "Bayesian fusion in cancer gene prediction." *Int J Comput Appl* 1 (2014): 5-10.

[20] de Trad, Chafia Hejase, Qiang Fang, and Irena Cosic. "Protein sequence comparison based on the wavelet transform approach." *Protein engineering* 15.3 (2002): 193-203.

[21] Akhil, M., B. L. Deekshatulu, and P. Chandra. "ScienceDirect International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013 Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm." *Procedia Technol* 10 (2013): 85-94.

[22] Wassfy HM, Abd Elnaby MM, Salem ML, et al. Eukaryotic gene prediction using advanced DNA numerical representation schemes. In: *Processing of Fifth International Conference Advances in Applied Science and Environmental Engineering (ASEE), Kuala Lumpur, Malaysia*, 2016.

[23] Zelyas, Nathan, et al. "Precision response to the rise of the SARS-CoV-2 B. 1.1. 7 variant of concern by combining novel PCR assays and genome sequencing for rapid variant detection and surveillance." *Microbiology spectrum* 9.1 (2021): e00315-21.

[24] Chung, Hsing-Yi, et al. "Emergency SARS-CoV-2 variants of concern: novel multiplex real-time RT-PCR assay for rapid detection and surveillance." *Microbiology Spectrum* 10.1 (2022): e02513-21.

[25] Erster, Oran, et al. "Rapid and high-throughput reverse transcriptase quantitative PCR (RT-qPCR) assay for identification and differentiation between SARS-CoV-2 variants B. 1.1. 7 and B. 1.351." *Microbiology spectrum* 9.2 (2021): e00506-21.

[26] Gravagnuolo, Alfredo Maria, et al. "High throughput diagnostics and dynamic risk assessment of SARS-CoV-2 variants of concern." *EBioMedicine* 70 (2021): 103540.

[27] Banada, Padmapriya, et al. "A simple reverse transcriptase PCR melting-temperature assay to rapidly screen for widely circulating SARS-CoV-2 variants." *Journal of clinical microbiology* 59.10 (2021): e00845-21.

[28] Wang, Huanyu, et al. "Mutation-specific SARS-CoV-2 PCR screen: rapid and accurate detection of variants of concern and the identification of a newly emerging variant with spike L452R mutation." *Journal of clinical microbiology* 59.8 (2021): e00926-21.

APPENDIX

```
clc;

clear all;

%close all;

%clf;

tic

%seq_file = fopen("C:\Users\Anjali Rana\OneDrive\Desktop\bq1.txt", 'r');

%a = fscanf(seq_file, '%c');

%a = regexprep(a, '[\s\n]+', '');

x=textread ('OX430870.1.dat','%s');

a=char(x);

[r2 c2]=size(a);

c=0;

for i1=1:r2

for j1=1:c2

c=[c a(i1,j1)];

end

end

% 2 converts the DNA characters in to numerical form

D_length=length(c);

for i=1:D_length-1

code=c(i+1);

switch (code)

case {'A','a'}

I5(:,i)=[0.1260]';

%I5(:,i)=[1]';

case {'G','g'}

I5(:,i)=[0.0806]';

%I5(:,i)=[2]';

case {'C','c'}
```

```
I5(:,I)=[0.1340]';
%I5(:,I)=[3]';
CASE {'T','T'}
I5(:,I)=[0.1335]';
%I5(:,I)=[4]';
END
END
%ABOVE CODE WILL GIVE YOU NUMERICAL SIGNAL
SD1=I5-MEAN(I5);
PLOT(SD1);
L_WIN=512;%64%242%220%82%50%80%50%50%30%351;%INPUT('ENTER THE LENGTH OF  WINDOW
FUNCTION:');M65145
N=1:L_WIN;
W(N)=RECTWIN(L_WIN);
%ZERO PADDING
MAZ = [ZEROS(1,(L_WIN)/2) SD1 ZEROS(1,(L_WIN)/2)];
S1=LENGTH(MAZ);
SAK1 =  ZEROS(29846,513);
    FOR I=1:S1-(L_WIN-1)
      M=MAZ(I:I+(L_WIN-1));
      UA1=M.*W(N);
%          UA1=XCORR(UA1);
%         SAK=PBURG(UA1,O_AR,N);
        SAK=PBURG(UA1,4,1024);
%          SAK=FFT(UA1,1024);
      SAK1(I,:)=SAK;
    END

    SD2=(ABS(SAK1)');
    SD3=SD2(1:512,:);
  SD3=(SD3/MAX(MAX(SD3)));
```

```
imagesc(sd3)
xlabel('Nucleotide position')
ylabel('Frequency bins')
colorbar


sd4=sum(sum(sd3.^4));
    sd5=sum(sum(sd3.^2)).^2;
     sd6=sd4/sd5;
     disp(sd6)
Threshold = 7.46*10^-8;
 if sd6>Threshold
    name = 'covid negative';
    X = [name];
    disp(X)
        else
    name = 'covid positive';
   X = [name];
    disp(X)
 end
```