ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

Classification and Alignment of SARS-Coronavirus Sequences: A Machine Learning and Bioinformatics Approach for Drug Discovery

Sweeti Sah[1], B. Surendiran[2], R. Dhanalakshmi[3], Akash Kamerkar[4]
[1,2,3,4]Department of Computer Science and Engineering
[1,2,4]National Institute of Technology Puducherry, Karaikal, India
[3]IIIT Tiruchirappalli, Trichy, India
Email: [1]sweetisah3@gmail.com, [2]surendiran@nitpy.ac.in, [3]r_dhanalakshmi@yahoo.com, [4]akash.kamerkar@gmail.com.

***Abstract*---Today Genomics is the most powerful field of research bounding several other technologies like deep learning, computer vision, machine learning and natural language processing. As SARS-Coronavirus is a global pandemic disease and is growing exponentially all across the world, classification helps the virologist to differentiate the viruses through their sequences in several pre-defined classes and further collecting the amino acid sequences of the classified virus to understand the homology between the protein sequences. The alignment of the common protein sequence is done for classified SARS-Coronavirus, as one virus can have many protein sequences. Drugs that are developed for curing the disease target basically protein. Understanding the protein sequence helps in discovering drugs for curing the disease. This research work consists of classification of nucleotide sequences of SARS-Coronavirus from other viruses like MERS, hCOV-19, and other viruses using machine learning classifiers like naïve bayes, random forest, decision tree and KNN classifier. As it can able to generate information for diagnosing the disease at the biomolecular level and can open up unlimited doors for curing and treating the disease at a most basic level. From the dataset of 1055 sequences, we classified the viruses according to their class labels and validated them using 5-fold cross-validation. The algorithms show the best result in ngram (6,6) with a maximum accuracy of 99.8%. Finally, collecting the protein sequences of classified SARS-Coronavirus from the NCBI website and aligning the amino acid sequences to observe the similarity percentage among them, which can further help to classify the protein families.**

Key words: SARS-Coronavirus, Nucleotide Sequences, Amino Acid Sequences, Classification Algorithm, Machine Learning, Alignment.

I. INTRODUCTION

Over the decades, the issue of virus categorization is always a concern for epidemiology or virology. The major challenge is to identify the detection of unknown viruses. In the present situation Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is mutating very fast and creating divergent variants, hence creating the prediction of the virus even more challenging. Apart from novel coronavirus, two more coronaviruses came into existence that is MERS-CoV and SARS-CoV-1. Machine learning plays a major role in predicting the coronaviruses by understanding the divergent genetic functional characteristics [1]. Genomic Structure analysis by sequencing is presently the ongoing topic to understand molecular dynamics [2]. Change in genomic structure causes properties to change and also result from difficulty in finding the drug for the treatment of disease [3]. The list of Pandemics till now are given below with their total deaths and time duration in Table 1, [4].

TABLE I: VARIOUS PANDEMIC OVER TIME

Pandemics	Time Period	Deaths
Black death	1347-1351	200 Million
New World Smallpox Outbreak	1520-onwards	56 Million
Spanish Flu	1918-1919	40-50 Million
HIV/AIDS	1981-Present	25-35 Million
Cholera Pandemics 1-6	1817-1923	1 Million+

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

Swine Flu	2009-2010	200,000
Ebola	2014-2016	11,000
MERS	2015-Present	850
SARS	2002-2003	770
COVID-19	2019-Present	778,228 as on Aug 18, 2020

The Coronavirus Outbreak has created a startling situation at the national and international levels. Researchers all over the world are now working to battle with COVID-19 to avoid the spread of COVID-19 all across the world [5]. COVID-19 has become a global pandemic disease affecting more than 100 000 people in 100 countries. The present outbreak situation is very much alike and non-identical to a prior severe acute respiratory syndrome that is MERS, SARS and etc. [6]. Coronaviruses are RNA viruses that are single-stranded positive-sense which contain the largest viral genomes of length up to 32kbps. The family Coronaviride contain four genera that is Alphacoronavirus and Betacoronavirus can contaminate mammalian hosts while the other two Gammacoronavirus and Deltacoronavirus infect avian hosts. Coronavirus genomes consist of genomic plasticity like other RNA viruses [7]. **Fig 1** shows the single stranded RNA structure.

Fig. 1. RNA Single Stranded Structure

RNA virus has potential of mutating in the human body. To assimilate the evolution of this virus, precise determination of mutation rate is required in order to show the risk of infectious disease [8]. Spike (S), Membrane (M), Nucleocapsid (N), and Envelope (E) protein, these proteins play an essential role in the structure of the virus as well as in the replication cycle [9]. These proteins can be seen in **Fig 2,**

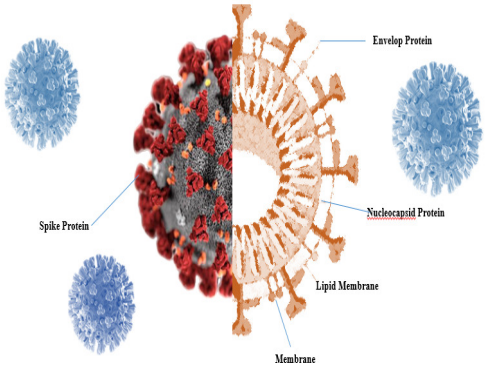

Fig. 2. SARS-CoV2 Structure

In order to detect positive cases of COVID-19 patient, pathological tests are being done which is very time consuming as well as centres and the number of kits are also limited. 60% of patients surprisingly getting SARS-CoV-2 positive even they are not having any symptoms. The major issue is that no vaccine is yet discovered. With the help of machine learning techniques [5], many clinical outcomes can be predicted like data related to patient medical history, geographical location and symptoms of the patient. The technique is used to detect the virus existence which is used especially by doctors and practitioners as a tool for decision making.

The organization of the paper is as follows, Section I consists of Introduction, in Section II Literature reviews of previous work are mentioned. Section III and Section IV describe Methodology and Experimental Analysis of various classification algorithms (Multinomial Naïve Bayes Algorithm, Random Forest Algorithm, KNN Algorithm and Decision Tree Algorithm) and Protein Sequence Alignment of Classified SARS-Coronaviruses to compute the similarity. Section V shows the result of the proposed work. Section VI describes its Conclusion. Finally, Section VII shows the future scope.

I. LITERATURE REVIEW

The classification approach using machine learning helped to separate the SARS-Coronavirus nucleotide

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

sequence. As protein sequence alignment can be beneficial in order to find the similarity between the sequences. So, the classified sequences were aligned as it can able to generate information for diagnosing the disease at the biomolecular level. Decoding the DNA, RNA and Protein sequence can open up unlimited doors for curing and treating the disease at a most basic level. The drugs target protein so it's necessary to understand the protein functionality. Several research works have been carried out with the genomic dataset to understand and classify the SARS-Coronavirus sequences. This approach may help in effective drug discovery. Few research articles have been found in the survey regarding the SARS-Coronavirus sequences, which is discussed below. The literature survey is broadly grouped into the various approaches like classification and prediction. The classification of CT images using DeepSense Model [10], and prediction includes the cases prediction using various models like heterogenous ensemble forecasting model [11], prediction of mortality rate and the risk associated with it [12] etc. [1] proposed a machine learning technique which is called COVID-Predictor, in which 1000 sequences of SARS-CoV-1, SAR-CoV-2, MERS-CoV and various other viruses are taken into consideration to training the viruses using Naïve Bayes Classifier. This helps in predicting the unknown sequence of these viruses. K-mer and n-gram techniques are generated in order to develop COVID-Predictor and construction of feature vector by a descriptor of the sequence. The validation done is 10-fold cross-validation in comparison with other techniques of classification. The result shows that it is able to achieve 99.3% average accuracy on the unseen validation sets. They have designed a web-based application for the same pre-trained model where the unknown sequence of viruses can predict the class of coronavirus by uploading. Similarly, [13] proposed a model to distinguish 9238 sequences in three stages that is pre-processing of data, labelling and classification of data. Pre-processing of data transforms COVID-19 amino acids into eight clusters of numbers on the basis of data labelling of 27 countries that is from 0 to 26. The initial method consists of a choice of one number by code number for each country. The next method is based on only binary elements for each country. For discovering different COVID-19 protein sequences Classification algorithms have been executed country wise. The result showed 100% accuracy for country-based binary labelling method with classification algorithms like Linear Regression, Support Vector Machine, K-Nearest Neighbour classifiers. Finally, the USA with huge information in infection rate has a high preference for the right classification as compared to other countries with less information rate. The unequal data protein sequences of COVID-19 are a major issue. In the USA the 76% of data expressed from total of 9238 sequences.

Another machine learning approach have been developed using the peculiar genomic signature of novel pathogens classification to give rapid alignment-free taxonomic. [7] This paper showed an accurate classification of the COVID-19 virus. The major contribution includes, first discovering intrinsic viral genomic signature for real-time and more authentic machine learning classification of COVID-19 pathogen sequences. The second approach is bare-bones of DNA sequences and didn't require gene or genome annotation. The third is a successive refinement of classification which is the decision tree approach. Fourth is Spearman's rank correlation analysis to approve result and COVID-19 sequences to the familiar genera (family Coronaviridae) and sub-genera (genus Betacoronavirus). [14] Analysed hCov genome sequences by clustering the sequence on the basis of geographical location and constructing the phylogenetic tree to understand its evolutionary history among the countries then classifying the virus to identify the virulence of the strains of many countries either as severe or mild and extraction of features from its genetic material. Finally, predicting the mutation using deep learning models. The pipelined analysis of hCov genome sequence shows results consisting of sequences of 67 countries, which was collected from the GISAID dataset. Identification and classification of genes are useful in understanding the function of new protein. The authors of [15] has given a review in order to classify the infected and normal genes with the help of machine learning techniques. The study includes more details related to bioinformatics as well as key issues in DNA Sequencing using machine learning approach.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

[8] Illustrated mutation rate of complete genomic sequence that is collected from the patient's dataset of various countries. Then, to regulate the nucleotide mutation and Condon mutation separately the dataset is processed. According to the size of the dataset, the mutation rate is divided into four zones that is China, Australia, the United States and the Rest of the World. The more amount of thymine (T) and adenine (A) are generally mutated to other nucleotides for all zone but on the other hand, the codon is not mutating frequently like nucleotides. In order to forecast the future mutation rate of the virus, a recurrent neural network model has been used. This model gives a Root Mean Square Error of 0.06 during testing and while training it gave 0.04 which is considered as an optimized value. With the help of the training and testing process, the mutation rate of nucleotide is predicted for 400th patient. By mutating nucleotides from T to C and G, C to G and G to T, only 0.1% increment is found in mutation rate. While 0.1% decrement is noticed for mutating of T to A and A to C. Hence this model is able to forecast day basis mutation rates if huge data of a patient is present. This paper has not focused on predicting the protein structure of the genomic sequence. So, in order to predict the protein 3D structure of protein structure, Alpha Fold Algorithm have been proposed. [16] this has been able to predict protein structures of COVID-19. If given amino acids sequence, which is an essential block of protein, the AlphaFold algorithm has the potential to predict the 3D structure of the protein, although this is an intensive process that needs many protein visualization techniques and structural analysis like nuclear magnetic resonance, cryo-electron microscopy and X-ray crystallography. This algorithm consists of three different layers of deep neural networks. It conducted training on the Protein Data Bank (Open-Source Database). This database consists of 3D structures for huge biological molecules like nucleic acids and proteins. The output shows distogram that consists of predicted secondary structure and accessible surface area. These proteins include protein 3a, membrane protein, nsp4, nsp6, nsp2 and papain-like C terminal domain.

II. METHODOLOGY

A. Dataset Description

This dataset has 1055 Sequences of various viruses all across the world like SARS-Coronavirus, MERS, hCOV-19 and other viruses. The first column of the dataset describes PID which is the unique accession number of the virus-like KY426711, which can be found on the NCBI website easily. Then the second column shows the nucleotide sequence of the virus. Finally, the third column shows the various classes of the virus. As shown in **Table 2**. This dataset information is available on the NCBI website. [17]

Class 1: SARS Coronavirus

Class 2: MERS

Class 3: hCoV-19

Class 4: other viruses

TABLE II: DATASET ATTRIBUTES EXAMPLE

PID	SEQ	CLASS
KY426711 |Homo sapiens|Sierra Leone|2015	CGGACACACAAAAAGA AAGAAGAATTTTTAGGA TCTTTTGTGTGCG...	4
KF600634 |Homo sapiens|Saudi Arabia|2013/05/30	CTTGCAGAACTTTGATTT TAACGAACTTAAATAAA AGCCCTGTTGT...	2
AY502929 |SARS coronavirus TW6| complete genome	ATATTAGGTTTTTACCTA CCCAGGAAAAGCCAACC AACCTCGATCT...	1
MF741830 |Homo sapiens|Jordan|2015/08/31	ATCTCACTTCCCCTCGTT CTCTTGCAGAACTTTGAT TTTAACGAAC...	2
KC762622 |Homo sapiens|Indonesia|2008/04/09	AGTTGTTAGTCTACGTG GACCGACAAGAACAGTT TCGAATCGGAAG...	4

B. Proposed Workflow

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

Most of the literature works had analysed/classified Genomic Sequences. In this work, we had done alignment of classified SARS-Coronavirus by taking their amino acid sequences from NCBI website to find the average similarity between the sequences. The proposed methodology consists of classifying the sequences of the virus using machine learning classifiers and alignment of amino acid sequences to observe the changes and similarity between the sequences (Pairwise) has been mentioned in **Fig 3,**

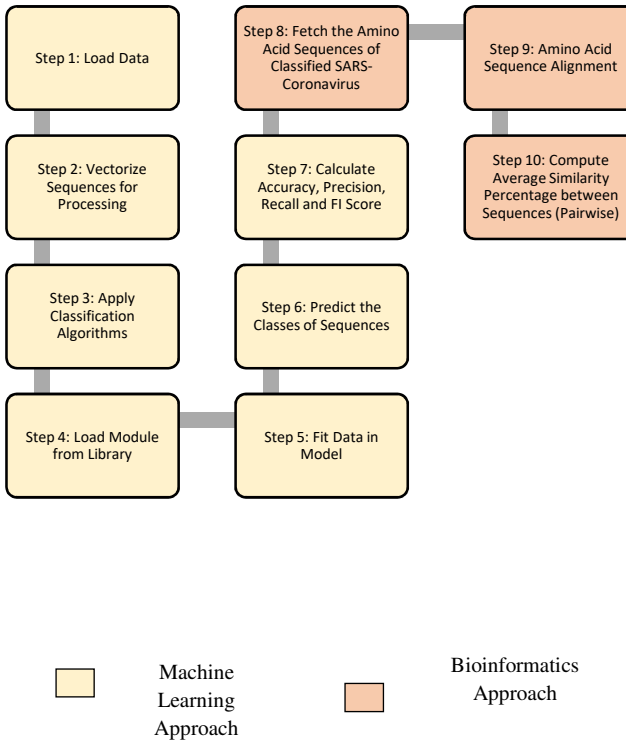

Fig. 3. Workflow of Proposed Work

Step 1: Import necessary libraries and loading data.

Step 2: Reading the genomic dataset. Here, **Fig 4** shows the sample view of nucleotide count for any particular genomic sequence. This step vectorizes the genomic dataset for processing further. We then create a function to convert strings into K-mer words. Here the value of K-mer is 6 and convert those sequences into a bag of words. Importing count vectorizer. Assigning each K-mer a vector value.

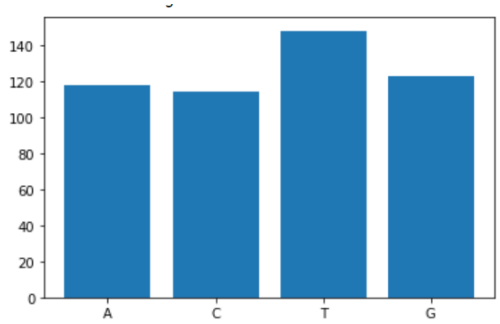

Fig. 4. Graphical View of Sequence Count for the particular sequence (Sample)

Step 3: Applied various machine learning classifier to the sequence to differentiate the sequences belonging to particular class.

Step 4: To apply classifiers, loaded necessary modules from library.

Step 5: Fit the genomic dataset into the model.

Step 6: Predict the classes of the nucleotide sequences. Splitting the dataset into a training set (70%) and a testing set (30%).

Step 7: Computed accuracy, precision, recall and F1 Score.

Step 8: Fetch the protein sequences of classified SARS-Coronavirus sequences from the NCBI website

Step 9: Pairwise Protein Sequence alignment (amino acid sequences)

Step 10: Computed average similarity percentage between the amino acid sequences to understand the protein sequence functionality.

C. Classification Algorithms

Here we are presenting the list of classifiers that are used in our work. A classifier is used to distinguish different objects based on some feature. It can be of many types. Few classification algorithms have been discussed in this the paper below in terms of

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

classifying the sequence of the various virus all across the world. [18]

1. **Multinomial Naïve Bayes Classifier:** This is based upon Bayes Theorem and is a probabilistic model which is easy to build, useful for the huge dataset and is given by the formula, [19]

$$P\left(\frac{B}{D}\right) = \left(\frac{P\left(\frac{D}{B}\right) * P(B)}{P(D)}\right) \quad (1)$$

In this, we are finding the probability of B that is happening, given that D, has already occurred. Where D indicates evidence and B specifies hypothesis. Here consideration is that, all features are independent and the existence of one feature will not affect others, this is called naïve. [18] or we can say in other terms it is a conditional probability [19], where

$P\left(\frac{B}{D}\right)$ = Posterior Probability

$P\left(\frac{D}{B}\right)$ = Likelihood

$P(B)$ = Class prior probability

$P(D)$ = Predictor prior probability

Now, for more features the conditional probability is given by, [18]

$$P\left(\frac{Z}{x1,x2,\ldots,xn}\right) = \left(P\left(\frac{x1}{Z}\right) P\left(\frac{x2}{Z}\right) P\left(\frac{x3}{Z}\right) \ldots\ldots P\left(\frac{xn}{Z}\right)\right) / P(x1)P(x2)P(x3)..P(xn) \quad (2)$$

Here $x1$, $x2$ …, xn represents features and variable z represents a class variable which represents whether it is suitable for a given condition or not.

$$P(z/x1, x2, \ldots., x3) \text{ Is directly proportional to } P(z)\ \Pi\ P\left(\frac{xi}{Z}\right) \quad (3)$$

The variable Z is the outcome and there could be a condition that classification can also be multivariate. So, we need to discover class z with maximum probability. [18]

$$Z = \text{argmax}_z\ P(z)\ \Pi\ P\left(\frac{xi}{z}\right) \quad (4)$$

Here we need to consider argmax_z in order to get the highest probability.

2. **Random Forest Classifier:** This is a supervised algorithm, used for regression and classification. It consists of a tree, more trees mean more robust a forest is. It is considered more accurate and robust because of the number of decision trees taking part in the process. There is no overfitting problem and can handle missing values too but the random forest is a little slow in generating predictions as it is having multiple decision trees and is also hard to interpret as compared to the decision tree. [20]

3. **Decision Tree Classifier:** This is a supervised machine learning algorithm. It resembles like flowchart arrangement in which internal nodes shows test on the feature, leaf shows the class label, and branches show conjunctions of features to those of class labels. Finally, the classification rule means the path from the root to leaf. [21]

4. **KNNs Classifier:** KNN is a supervised learning algorithm that uses labelled input information set to forecast the data points of output. It is a very simple algorithm that can be used for a different set of problems. KNN has feature similarity means it checks how identical a data point is to the near neighbour and classifies it according to the most identical data points. KNN does not make any presumption about the data set before which makes the algorithm more attractive as it can handle realistic data. This is a lazy algorithm as it retains the training data set instead of learning the data set. It can be used for both classification and regression problems. [22]

D. Protein Sequence Alignment

The classified 149 sequences of SARS-Coronavirus, we collected amino acid sequences for all the

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

classified SARS-Coronavirus, out of which, we took the few common proteins (collected from the NCBI website) and compared them together using alignment tool and found the similarity between the amino acid sequences of several SARS-coronaviruses. Various proteins and its function are shown below in table 3 below, [23]

TABLE III: PROTEIN STRUCTURE AND ITS DESCRIPTION

Protein	**Description**
E protein	It interrelates with M protein to shape E protein
M protein	This is the Central supervisor of CoV assembly and also tell viral envelope shape
N protein	This ties up with an RNA genome to create nucleocapsid
S protein	It is responsible for entry to host cell by holding up with host cell receptors

IV EXPERIMENTAL ANALYSIS

Various Performance Metrics are given below, [24]

Predicted \ Actual		
	TrPo	FaPo
	FaNe	TrNe

Confusion Matrix

Accuracy: This is a good measure in which target variable class in the data are mostly balanced. Where TrPo is true positive, TrNe is true negative, FaNe is false negative and FrNe is false negative.

$$\text{Accuracy (A)} = \frac{TrPo+TrNe}{TrPo+FaPo+FaNe+TrNe}$$

Precision: This is positive predictive value that refers to the fraction of relevant instance among the total instances retrieved.

$$\text{Precision (P)} = \frac{TrPo}{TrPo+FaPo}$$

Recall or Sensitivity: Sensitivity is fraction of applicable instances retrieved over the entire relevant instances.

$$\text{Recall (R)} = \frac{TrPo}{TrPo+FaNe}$$

F1 Score: F1 Score is the symphonic average of recall and precision measurement. Value 1 shows perfect F1 Score and value 0 shows worst F1 Score.

$$\text{F1 Score} = 2 * P * R/(P + R)$$

TABLE IV: PERFORMANCE METRICS OF SEQUENCE CLASSIFICATION

Classifier	**Multinomial Naïve Bayes**		**Random Forest**		**Decision Tree**		**KNN**	
ngram	**(6,6)**	**(9,9)**	**(6,6)**	**(9,9)**	**(6,6)**	**(9,9)**	**(6,6)**	**(9,9)**
Accuracy	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.5 %
Precision	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.5 %
Recall	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.5 %
F1 Score	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.8 %	99.5 %

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

In **Table 4**, ngram is the set of occurrences of a continuous sequence of n length from a large set of sequence. The dataset is divided into 7:3 ratio means 70% of the data is used for training and 30% used for testing. The model is validated using 5-fold cross-validation. As a result, it is observed that multinomial naïve Bayes, random forest and decision tree classification algorithms show the same result with the same ngram whereas KNN shows different accuracy in different ngram. So, all the algorithms show the best result in ngram (6,6) with an accuracy of 99.8%.

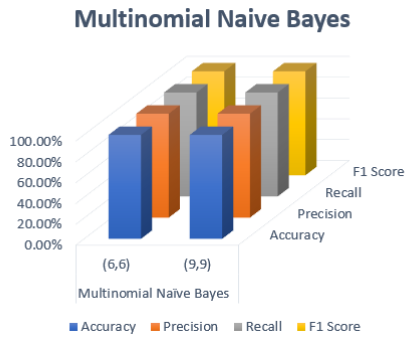

Fig. 5. Multinomial Naïve Bayes Performance Metrics with ngram (6,6) and (9,9)

Fig 5, shows the performance metric graph for Multinomial Naïve Bayes Classifier. Multinomial is used for text classification means whether the text belongs to the particular category or not. The features used by the classifier are the count of the occurrences of text in the sequence. [18]

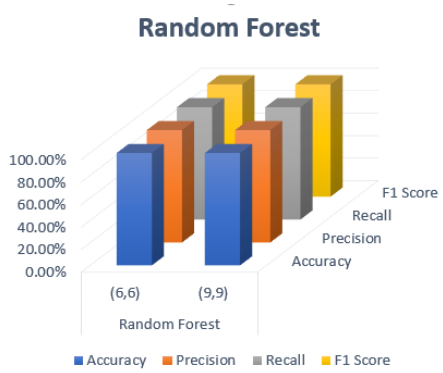

Fig. 6. Random Forest Performance Metrics with ngram (6,6) and (9,9)

Fig 6, shows the performance metric graph for Random Forest Classifier. This classification algorithm consists of many decision trees. It uses feature randomness and bagging; it creates uncorrelated forest consists of trees whose forecast is more accurate than any individual tree. Hence in a random forest, we conclude with the trees that are not only trained with different datasets but also make use of different features to make decisions. [25]

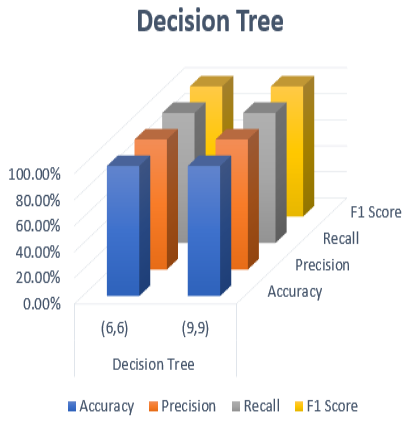

Fig. 7. Decision Tree Performance Metrics with ngram (6,6) and (9,9)

Fig 7, shows the performance metric graph for Decision Classifier. Decision Tree Classifier is very simple to understand, visualize and interpret. It can handle data like numerical, categorical also can handle multi-output problems. It implicitly performs feature selection and requires less effort for data preparation from users. Hence the nonlinear relationships do not alter tree performance between parameters. [26]

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

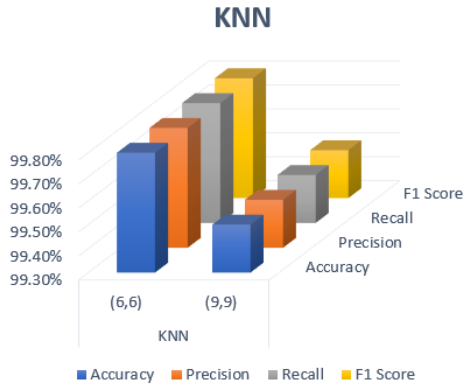

Fig 8: KNN Performance Metrics with ngram (6,6) and (9,9)

Fig 8, shows the performance metric graph for KNN Classifier. KNN Classifier is a non-parametric classifier that is performed by using a training dataset means input and target variables and comparing the test data. [27]. After classification of SARS-Coronavirus from other viruses, there amino acid sequences were aligned and finally found average similarity of Amino Acid Sequences. The similarity percentage was computed using the python programming language by comparing the sequences pairwise. The results may vary with the change in the amino acid sequences. The results are based on the dataset which we have used. The result only consists of a few amino acid sequences after alignment and comparing pairwise to find the similarity between them.

V. RESULT

This research work consists of the classification of SARS-Coronavirus sequences from other viruses like MERS, hCOV-19 and other viruses. [1] proposed the classification of SARS-Coronavirus with other using naïve bayes classifier as an existing work. The result of our work comprises of various classification algorithms that shows the accuracy of multinomial, random forest and decision tree is the same, whereas KNN is slightly different from others. We computed the accuracy of various classification algorithms at different ngram like ngram(6,6) and ngram(9,9) as shown above in **Table 4**. The essential concept, ngram is the set of occurrences of a contiguous sequence of n items from a large set of sequence. Hence our proposed work is able to classify the nucleotide sequence of various viruses with maximum accuracy but the accuracy might vary if the dataset increases. The dataset comprises viruses of SARS-Coronavirus, MERS, hCOV-19 and other viruses. After the classification of sequences, obtained the amino acid sequences of viruses and found the similarity percentage between the sequences after doing alignment of amino acid sequences as shown in **Table 5.**

TABLE V: AVERAGE SIMILARITY PERCENTAGE OF AMINO ACID SEQUENCES (PAIRWISE)

Proteins	**Average Pairwise Similarity between Amino Acid Sequences of various SARS-Coronavirus**
Nucleocapsid protein	99.9 %
ORF1a	99.8 %
Polyprotein ORF1a	99.9 %
Putative Spike Glycoprotein	99.7 %
Replicase	99.8 %
Spike Glycoprotein	99.9 %

VI. CONCLUSION

As SARS-CoV-2 is a pandemic disease and is growing exponentially all across the world so the classification of the various virus from others helps the virologist to distinguish the virus through there sequences. As it can generate information for diagnosing the disease at the biomolecular level. we considered the sequences of various viruses and used machine learning approaches of classification. From the dataset of 1055 sequences, we successfully classified the viruses according to their class labels and validated them using 5-fold cross-validation. The algorithms showed the best result in ngram (6,6) with a maximum accuracy of 99.8%. Finally, collected the protein sequences of classified SARS-Coronavirus from the NCBI website and aligned the amino acid sequences to observe the similarity percentage among them. This work may help in effective drug discovery. Further work can include, the prediction of protein structure and understanding the mutation rate among them.

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

VII. FUTURE SCOPE

Future work in the field of machine learning with respect to SARS-CoV2 may include the following, [28]

1. Classification of treatments and their concentration followed by the prediction.
2. Detecting the drug to heal from SARS-CoV2.
3. Classifying the Human cell-based on treatment type and concentration using various deep learning approaches.

REFERENCES

1. Sarkar, Jnanendra Prasad, Indrajit Saha, Arijit Seal, and Debasree Maity. "COVID-Predictor: RNA Sequence based Prediction of Coronavirus." (2020).
2. Paden, Clinton R., Ying Tao, Krista Queen, Jing Zhang, Yan Li, Anna Uehara, and Suxiang Tong. "Rapid, sensitive, full-genome sequencing of severe acute respiratory syndrome coronavirus 2." Emerging infectious diseases 26, no. 10 (2020): 2401.
3. Lan, Jun, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang et al. "Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor." Nature 581, no. 7807 (2020): 215-220.
4. Rekha Hanumanthu, Swapna. "Role of Intelligent Computing in COVID-19 Prognosis: A State-of-the-Art Review." Chaos, Solitons & Fractals (2020): 109947.
5. Bhonde, Swati, Madhulika Bhati, and Jayashree Prasad. "Predictive Analytics to Combat with COVID-19 using Genome Sequencing." Available at SSRN 3580692 (2020).
6. Wu, Zunyou, and Jennifer M. McGoogan. "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention." Jama 323, no. 13 (2020): 1239-1242.
7. Randhawa, Gurjit S., Maximillian PM Soltysiak, Hadi El Roz, Camila PE de Souza, Kathleen A. Hill, and Lila Kari. "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study." Plos one 15, no. 4 (2020): e0232391.
8. Pathan, Refat Khan, Munmun Biswas, and Mayeen Uddin Khandaker. "Time Series Prediction of COVID-19 by Mutation Rate Analysis using Recurrent Neural Network-based LSTM Model." Chaos, Solitons & Fractals (2020): 110018.
9. Schoeman, Dewald, and Burtram C. Fielding. "Coronavirus envelope protein: current knowledge." Virology journal 16, no. 1 (2019): 1-22.
10. Alaa O. Khadisos, Srihari Kannan, Yuvaraj Natarajan, Sachi Nandan Moahnty, & George Tsaramirsis , Analysis of COVID-19 Infections on a CT Images Using DeepSence Model, Adil Kha-didos, Frontiers in Public Health, (2020). Doi : doi.org/10.3389/fpubh.2020.599550.
11. Sharma, N., Dev, J., Mangla, M., Wadhwa, V. M., Mohanty, S. N., & Kakkar, D. A Heterogeneous Ensemble Forecasting Model for Disease Prediction. New Generation Computing,1-15.
12. Satpathy, S., Mangla, M., Sharma, N., Deshmukh, H., & Mohanty, S. (2021). Predicting mortality rate and associated risks in COVID-19 patients. Spatial Information Research, 1-10.
13. Alkady, Walaa, Muhammad Zanaty, and Heba M. Afify. "Computational Predictions for Protein Sequences of COVID-19 Virus via Machine Learning Algorithms." (2020).
14. Sawmya, Shashata, Arpita Saha, Sadia Tasnim, Naser Anjum, Md Toufikuzzaman, Ali Haisam Muhammad Rafid, Mohammad Saifur Rahman, and M. Sohel Rahman. "Analyzing hCov genome sequences: Applying Machine Intelligence and beyond." *bioRxiv* (2020).
15. Dixit, Pooja, and Ghanshyam I. Prajapati. "Machine learning in bioinformatics: A novel approach for dna sequencing." In 2015 Fifth International Conference on Advanced

ISSN (ONLINE): 2454-9762
ISSN (PRINT): 2454-9762
Available online at www.ijarmate.com

IJARMATE

International Journal of Advanced Research in Management, Architecture, Technology and Engineering (IJARMATE)
Vol. 7, Issue 5, May 2021

Computing & Communication Technologies, pp. 41-47. IEEE, 2015.
16. Faris Gulamali, "AlphaFold Algorithm Predicts COVID-19 Protein Structure", https://www.infoq.com /news/2020/03/deepmind-covid-19/, March 31, 2020.
17. National Center for Biotechnology Information (NCBI) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2017 Apr 06]. Available from: https://www.ncbi.nlm.nih.gov/
18. Rohith Gandhi, "Naïve Bayes Classification", https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c, May 5, 2018.
19. Sunil Ray, " Analytics Vidya", https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/, Sep 11, 2017
20. Avinash Navlani, "Understanding Random Forest Classifier in Python", https://www.datacamp.com/community/tutorials/random-forests-classifier-python, May 16, 2018.
21. Prince Yadav, "Decision Tree in machine learning", https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96, Nov 14, 2018.
22. Zulaikha Lateef, "KNN Algorithm: A Practical Implementation of KNN Algorithm in R", https://www.edureka.co/blog/knn-algorithm-in-r/, May 14, 2020.
23. Khan I, Ahmed Z, Sarwar A, et al., "The Potential Vaccine Component for COVID-19: A Comprehensive Review of Global Vaccine Development Efforts", Cureus 12(6): e8871. doi:10.7759/cureus.8871, June 27, 2020.
24. Mohammed Sunasra, "Performance Metrics for Classification problems in Machine Learning", Nov 11, 2017.
25. Tony Yiu, "Understanding Random Forest", https://towardsdatascience.com/understanding-random-forest-58381e0602d2, June 12, 2019.
26. Prashant Gupta, "Decision Tree in Machine Learning", https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052, May 18, 2017.
27. Ramteke, R. J., and Khachane Y. Monali. "Automatic medical image classification and abnormality detection using k-nearest neighbour." International Journal of Advanced Computer Research 2, no. 4 (2012): 190.
28. Khalifa, Nour Eldeen M., Mohamed Hamed N. Taha, Gunasekaran Manogaran, and Mohamed Loey. "A deep learning model and machine learning methods for the classification of potential coronavirus treatments on a single human cell." Journal of Nanoparticle Research 22, no. 11 (2020): 1-13.