VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2021, ISSN(e): 2308-8168, ISSN(p): 2411-6335
Volume 9, Number 1, January-December , 2021 pp:23-33

Discrimination of SARS-COV2 virus protein strain of three major affected countries: USA, China, and Germany

Dr. Khalid Allehaibi

[1]Department of Computer Sciences, Faculty of Computing and Information Technology, King Abdulaziz University Jeddah, Saudi Arabia
*Email: kallehaibi@kau.edu.sa

ABSTRACT

In this paper, we discuss the discrimination of SARS-COV2 viruses associated with three major affected countries the USA, China, and Germany. The discrimination can reveal the mutation as the result of viral transmission and its spread due to mutation associated with its protein structure which makes small changes in the Spike protein. To investigate the mutation in SARS-COV2, we downloaded the protein strains associated with the USA, China, and Germany from the UniProtKB by advance search through SARS-COV2, country name, and protein name: Accessory protein 7b, 6, ORF3a, 10, 8 protein, Envelope small membrane protein, Nucleoprotein, Membrane protein, Spike glycoprotein, 3C-like proteinase, and 2'-O-methyltransferase. After retrieving the protein sequences, we transform the biological form of sequences to their equivalent numerical form by using statistical moments. Further classification algorithms like Random Forest, SVM are used for their training and classification. Finally, performance evaluation is carried out using K-fold cross-validation, independent testing, self-consistency, and jackknife testing. The result received through all testing is more than 97%, which shows the visible discrimination among the protein strains of mentioned countries, which shows the strong mutation in SARS-Cov2 sequences.

KEYWORDS

SARS-COV2, Mutation, USA, Germany, China, Machine Learning, Statistical Moment

JOURNAL INFO

HISTORY: Received: October 15, 2021 Accepted:November 20, 2021 Published: December 2, 2021

INTRODUCTION

The World Health Organization declared the Severe acute respiratory Syndrome (SARS) as an international epidemic. The SARS was initially occurred in the Wuhun city of China in December 2019 [1][2][3] [4][5][6]. The SARS-COV2 name was given to the virus due to its resemblance with the existing virus SARS-COV that was reported for the first time in 2002. SARS-COV2. This causes respiratory disease also known as COVID-19. It is now a great threat to the human health throughout and the overall economy of the world. The disease was initially reported in the Wuhan city in China in December 2019 (Ren et al., 2020) [39]. In the very short time, the virus spread throughout the World. As of April 10, 2021 so far, the positive cases are 134,308,070 which brought the deaths of more than 2,907,944 people. The spread of the virus was very fast and affected almost all the countries in the World, USA and Germany are included in the most affected countries list as shown in the Table 1 as per the data retrieved up to April 2021, whereas China is included in this work as the virus is initiated so that to observe the possibility of mutation with respect to the china and further to USA and Germany as they are the most affected countries [9].

Table 1. Total Confirmed Cases and deaths reported up to April 10, 2021

Country	Cases	Deaths
United States	31,802,772	574,840
Germany	2,974,110	78,689
China	90000	4636

SARS-COV2 is closely similar to the old pathogenic corona viruses in human like SARS-COV which was spread in 2002 from china to various part of the World infected more than 8000 people which brought death of 774 people, Similarly MERS-COV was initiated in 2012 from the Middle East infected about 2494 people which caused the death of 858 with the very high fatality rate of 34.4%[40].

The SARS-COV2 genome is a single stranded positive-sense RNA (+ssRNA) is larger as compared to any viruses. The capsid outside the genome is formed by the nucleocapsid protein (N). Further the packing around the genome by structural proteins: spike protein (S), membrane protein (M) and and envelope protein (E) [7][8][10][41][42][50] . SARS-COV2 encodes various open reading frames (ORFs) like as ORF {1ab, 3a, 6, 7a, 8 and 10} located in three regions. These frames are predicted to code for the replicase polyprotein.

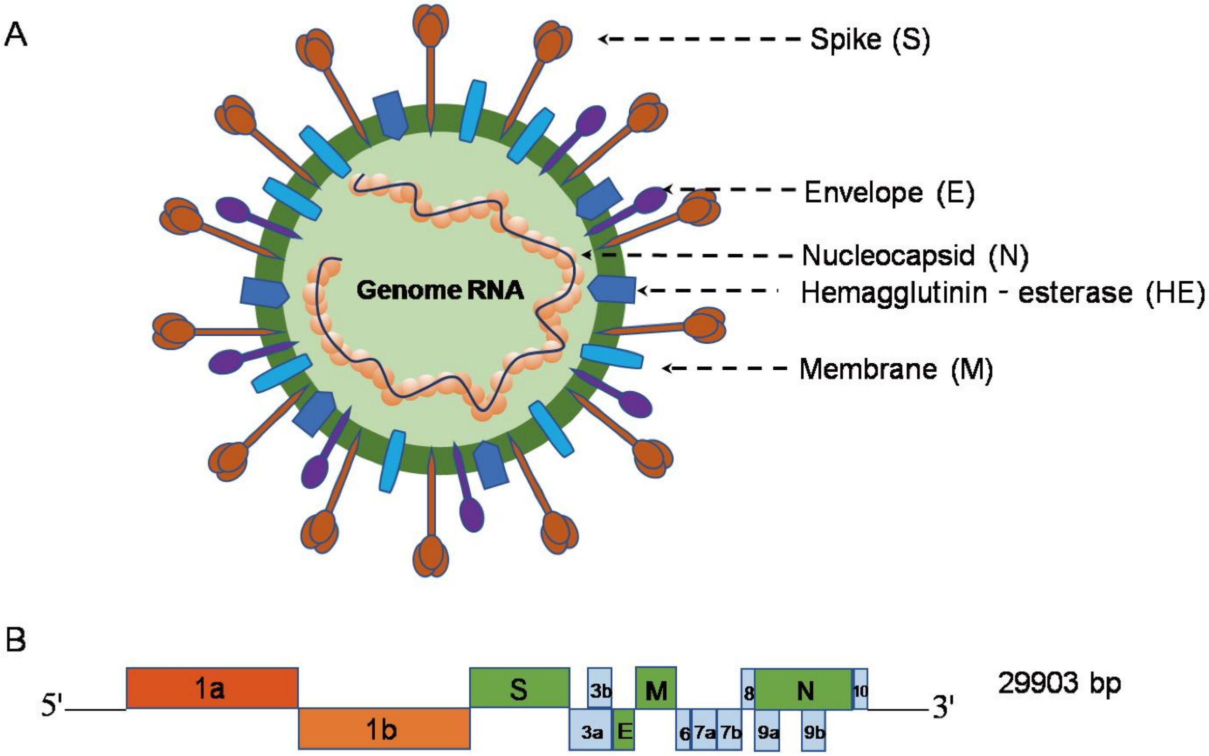


Figure 1. SARS-COV2 Structure representation [42][44]

It is observed through the existing studies that possible adaptation in amino acids composition and its structure heterogeneity in viral proteins is documented[11][12].. Therefore, Identification of SARS-COV2 viruses and further its discrimination, due to mutation, as compared to the same viruses in other part of world is very important to understand its mutation and further to understand its behavior which can be helpful for its prevention and its drug and vaccine discovery. The experimental identification of SARS-COV2 protein is not a simple process and an inexpensive task. The use of the latest technology from protein purification to sequence analysis by means of mass spectrometry after tryptic digestion, as described in [45], is quite time-consuming. Likewise, it is not realistic to analyze a large number of sequences from one part of the world compared to another part of the world in order to distinguish the pathogen or to identify place from where it is originated/mutated. Similarly, to identify ORF by homolog search and codon index technique has a very weak performance mainly due to the small dataset of verified sequences in the same way in the case of S, E N and M SARS-COV2 protein. Therefore, it is required to have the computational approaches in parallel to the traditional methods to identify these sites [13].

For the purpose various type of identifications approaches, different in nature, are used to identify SARS-COV2 virus as compared to other viruses or viruses in other animals, like the work done by Randhawa et al., they use taxonomic classification of the COVID-19 virus as *Sarbecovirus*, within *Betacoronavirus*, as well as quantitative evidence supporting a bat origin hypothesis [46]. They used linear SVM, Quadratic SVM Fine KNN and 10 fold cross validation for modeling and its performance evaluation

In the work, using genotyping, k-means clustering, time-evolution, sequence-alignment, algebraic topology, protein-folding stability and network theory, they classify the SARS-COV2 in US in four classes of substrains, they also identified that , 27964C>T-(S24L) on ORF8 has the highest mutation [47]. In the study [48], SARS-COV2 sequences of 44 outpatients and inpatients belong to the locals of the capital Montevideo of Uruguay in South America from March to May 2020. they observed in the genetic analysis the dominance of S and G clades which make 90% of the viral strains, they didn't get any correlated with the structural and non-structural protein. In another research, identification in the variation of N-protein of SARS-COV2 was discussed, for the purpose N-protein from the United state of America was campaered with the N-proteine sequences retrieved from the city Wuhan, China, they observed 107 mutations in N-protein [49].

In this study we developed a model based on the statistical moment for feature extraction and the classifiers Random Forest (RF) for its classification. For the model evaluation, we use 10-fold cross validation, Self-Consistency and Jackknife testing to calculate accuracy (Ac), sensitivity (Sn), specificity (Sp) and Mathews correlation coefficient (MCC). We obtained a very imprecise results which predict the possible adaptation in amino acids and structural heterogeneity in viral proteins is predicted.

MATERIALS AND METHODS: In this work four steps are used to develop and validate the machine learning model (1). Collection of data from the UniProt database (2). Conversion of Biological Sequences to the corresponding equivalent numerical form so that to use it for training and classification (3). In the third step we will use well-known classifiers to get the accurate classification so that to fine the discrimination in the data if exist. (4). Will use 10-Fold Cross validation, Jackknife testing, independent testing, and Self consistency to evaluate the performance of the classifiers as shown in the Figure 2.

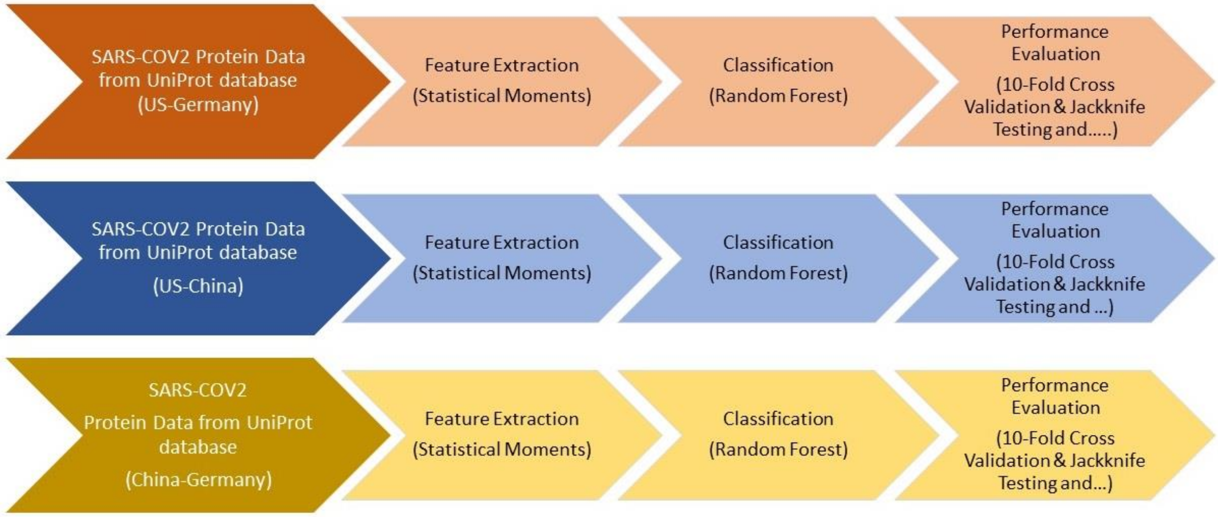

SARS-COV2 Protein Data from UniProt database (US-Germany)
Feature Extraction (Statistical Moments)
Classification (Random Forest)
Performance Evaluation (10-Fold Cross Validation & Jackknife Testing and.....)
SARS-COV2 Protein Data from UniProt database (US-China)
Feature Extraction (Statistical Moments)
Classification (Random Forest)
Performance Evaluation (10-Fold Cross Validation & Jackknife Testing and ...)
SARS-COV2 Protein Data from UniProt database (China-Germany)
Feature Extraction (Statistical Moments)
Classification (Random Forest)
Performance Evaluation (10-Fold Cross Validation & Jackknife Testing and...)

Figure 2. Methodology of the proposed work

BENCHMARK DATASET. In Machine learning problems, the valid dataset plays the most important role for the correct and valid decision. As the quality of predictor can be judge if the data is valid and downloaded from the authentic source. In this study, to investigate the mutation in SARS-COV2, we downloaded the protein strains associated with the USA, China, and Germany from the UniProtKB using advance search through SARS-COV2, country name, and protein name: Accessory protein 7b, 6, ORF3a, 10, 8 protein, Envelope small membrane protein, Nucleoprotein, Membrane protein, Spike glycoprotein, 3C-like proteinase, and 2'-O-methyltransferase.

FEATURE EXTRACTION. To classify the biological data, the feature extraction is an unavoidable activity, as without this we cannot classify the biological data. In this step the biological sequences are transform into vectors motifs, which are subsequence represent the attribute of the biological data. For such transformation many methods are developed like Aino Acid Composition AAC, Split Amino Acid Composition SAAC, Position Specific Storage Matrix PSSM, Tripeptide Compositing TPC, Pseudo-Amino Acid Composition PseAAC and statistical Moments [[17]-[30]]. In this work we will use the statical moment for the required feature extractions

STATISTICAL MOMENT: The proposed methodology develops on the use of Statistical moments to form a numerical representation such that the obscured information within the primary structure of proteins stays intact. These moments form a succinct numerical form such that the original data can be reconstructed without any significant loss of information. Moments can be obtained up to several orders, each provides a deeper perspective into specific aspects of data like positioning, eccentricity, skewness, and peculiarity [54]. Mathematicians and statisticians have devised many moments generating coefficients incarnated based on well-defined distribution functions and polynomials [[55]-[66]].

In the proposed work Hahn moments, raw moments, and central moments are organized to form a feature set. Hahn moment bear location and scale-oriented variance and are calculated based on Hahn polynomial. Central moments abide information regarding asymmetry, mean, and variance. The central moments are derived for the centroid of collective data making these moments scale variant and location invariant. Subsequently, raw moments are scale and location variant and represents properties like asymmetry, variance, and mean.

CLASSIFICATION: It is the supervised learning approach in which the computer learns from the given data which contains independent variables the features or attributes and the dependent variables the labels and makes the classification of the input data. The main aim is to identify the class of new input data will belong into,. In protein research, the popular classifiers are SVM, KNN and RF etc. In this project we restricted ourselves to SVM and RF to classify to train the model on the available dataset, we achieved best results through Random Forest so we discussed its outcomes.

1.1.1. Random Forest: It is a supervised ensemble learning algorithm used for both regression and classification. It constructs n decision trees for sub dataset randomly retrieved from the base dataset [34][35][36]. Further, every decision tree calculates the prediction result. At last, the final prediction is made based on the most voted result from the decision trees as shown in Figure 3.

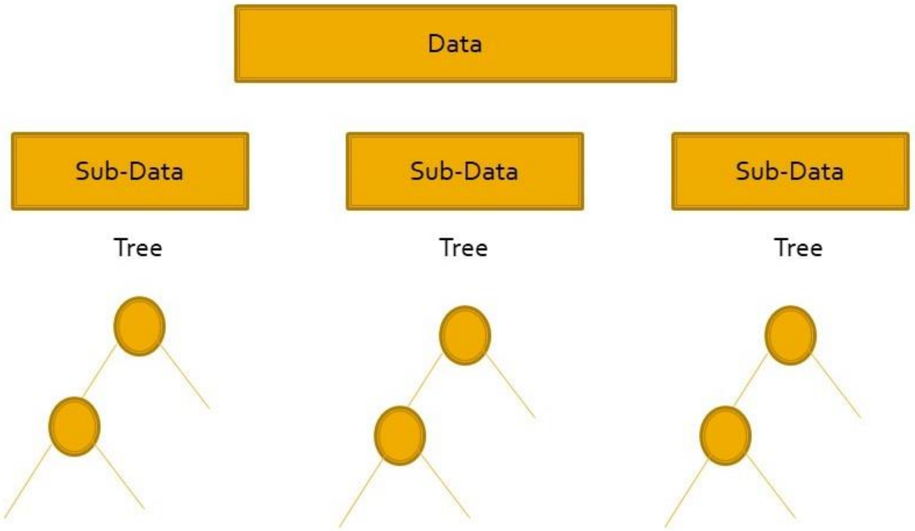


Figure 3 Random Forest Structure

1.2. Performance evaluation: Performance In this step, we check the performance of the classifier using the outcomes of the classifier based on the test data in the form of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Whereas TP is the number of actual positive labels which are predicted positive, TN is the number of actual negative labels which are predicted negative, FP is the number of negative labels which are predicted falsely positive, and FN is the number of positive labels which are predicted falsely negative labels, all are shown in the Table-1 which is called the confusion Matrix.

Table2 . Confusion Matrix.

Actual class	Predicted class	
	Negative	Positive
China Seq (Negative)	TN (predicted as China Seq)	FP (Predicted as Saudi Seq)
KSA Sequence (Positive)	FN (predicted as China Seq)	TP(Predicted as Saudi Seq)

Using the outcomes, TN, TP, FN and FP, the Accuracy Ac, Sensitivity Sn, Precession Pr, Specificity Sp and Matthew Correlation Coefficient MCC can be calculated using the following formulas:

- Accuracy: It is the percentage of accurate classification of the data set. It is the total number true positive and true negative out of the total number of cases.

$$\text{Ac} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (1)$$

- Sensitivity: This is the total number of positive correct classification. It is defined as

$$\text{Sn} = \frac{\text{TP}}{\text{TP+FN}} \quad (2)$$

- Precision: It is the proportion of correct positive classification (True positive) from cases that are predicted positive.

$$\text{Pr} = \frac{\text{TP}}{\text{TP+FN}} \quad (3)$$

- Specificity: It is the proportion of negative classification out of actual negative. It is defined as

$$:\text{Sp} = \frac{\text{TN}}{\text{FP+TN}} \quad (4)$$

- Mathews correlation coefficient (MCC): MCC is introduced by Brain W. Mathews, is a correlation coefficient use to predict and observe the classification test. It returns values from -1 to 1. If the value of the coefficient is 1, it represents the correct prediction and if 0, it indicates random prediction and if -1 then the prediction is not correct. The MCC is defined mathematically as:

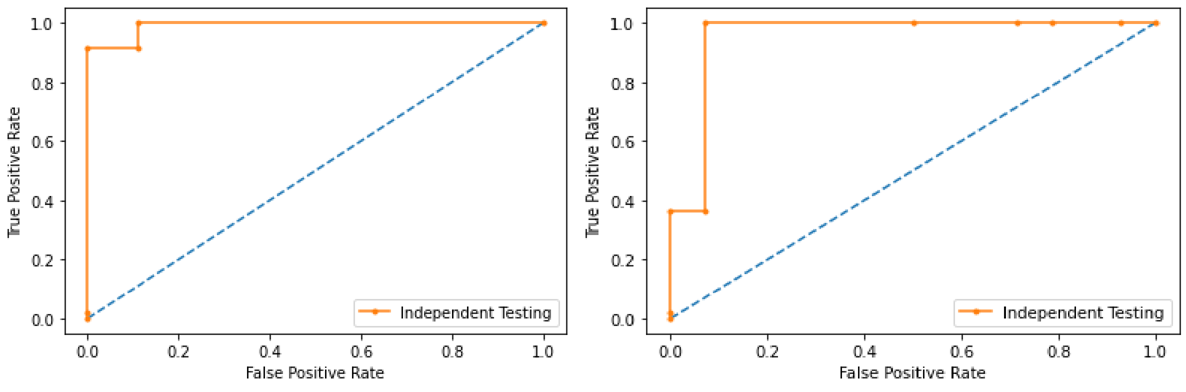
$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

Results and Discussion: To validate the performance of developed predictor, we will use the following three methods to evaluate the effectiveness of the model: Independent Testing, K-fold cross validation and the Jackknife Cross validation.

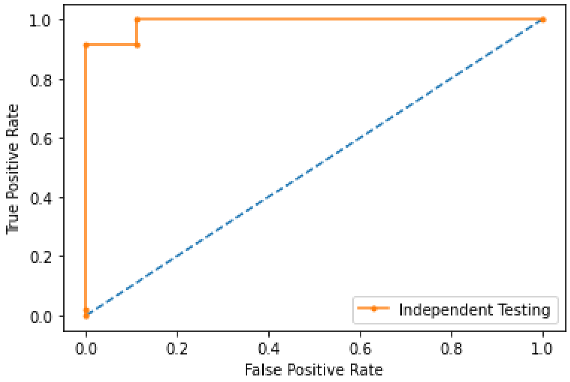
Independent Testing: This is one of the cross-validation methods which divides the dataset in two parts one for training and another for testing normally in 80:20 ratio. The predictor is trained by the training data using the selected classifier, in this work, we use the Random Forest, further the predictor is tested on the test data and calculate Ac, Sn, Pr, MCC and ROC. In this work we used three datasets one US to Germany, second is US to China and the third is China to Germany whose outcomes are presented in the Table 2 and by the ROC curves shown in the Figure 2, 3 and 4.

Table 3 Performance matrix using

Model	ACC (%)	MCC	SN (%)	Pr (%)	ROC
US-Germany	95.89	1.0	1.0	1.0	0.96
US-China	98.48	1.0	1.0	1.0	0.94
China-Germany	95.79	1.0	1.0	1.0	



(a). **(b)**



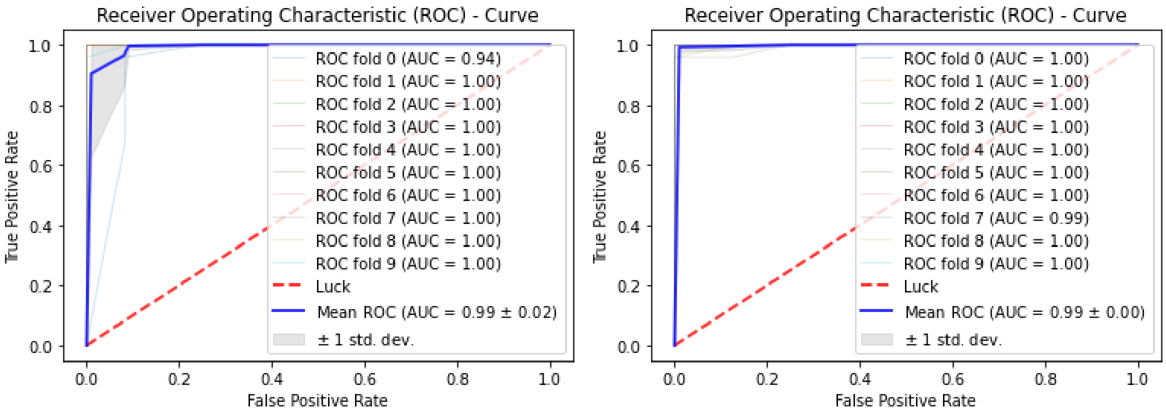
(c)

Figure 4 ROC Curve Independent Testing (a) US-Germany (b). US-China (c). China-Germany

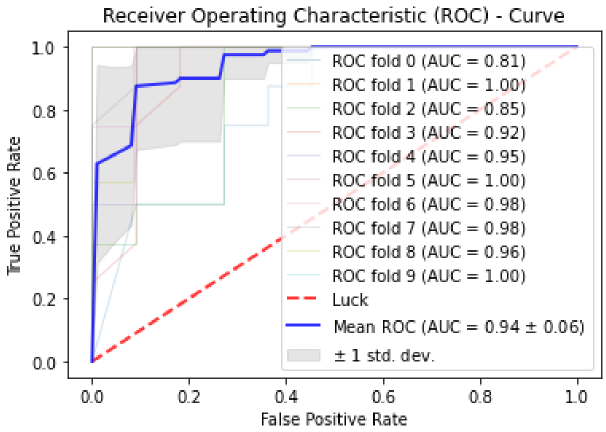
K-Fold cross validation. It is a very powerful and well-known method to evaluate the performance of the predictor. This method divides equally the data into K-folds, further it trains the model by K-1 folds and test by the remaining one-fold to estimate the accuracy Ac, Sensitivity Sn, Precession Pr and Matthew Correlation Coefficient MCC. The method repeats the process k-time fold by fold. Finally, the method calculates the average performance of all the K-folds calculated matrices. In this work we used K-fold Cross validation using K=10 for the tree datasets: US-Germany, US-China, and China-Germany

Table 4. 10 Fold Cross Validation Results

Model	ACC (%)	MCC	SN (%)	Pr (%)	ROC
US-Germany	97.22	90.91	0.9	1.0	0.99
US-China	100	1.0	1.0	1.0	0.99
China-Germany	95.79	1.0	1.0	1.0	



(a). (b)



(c)

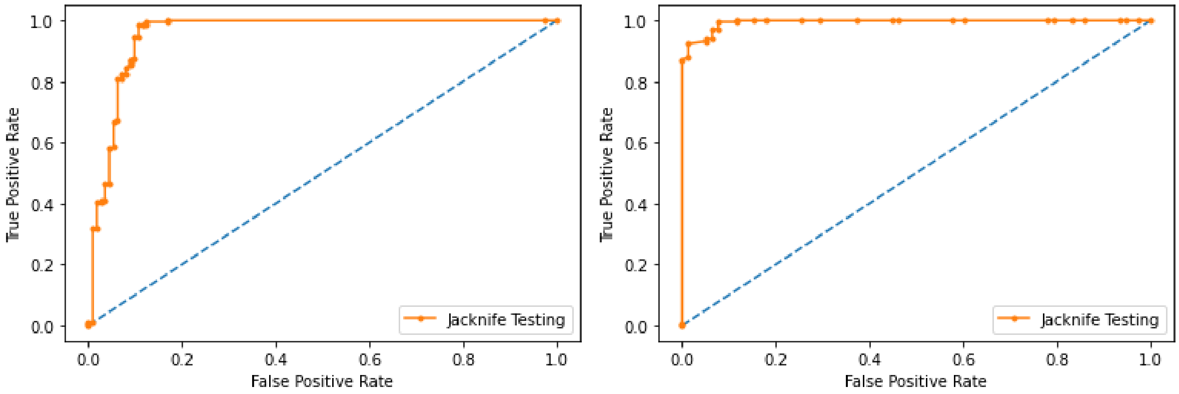
Figure 5. K-Fold Cross Validation (a). US-Germany (b). US-China (c). Germany-China

Jackknife Cross Validation:

This is the strong cross validation method used for evaluation the performance of the predictor. If the data contained N records, then the method train the model by N-1 records and test for the remaining one record therefore, this method is also called skip one strategy. The method repeats the process N times and make prediction for each sample data record. In this work the Jackknife test is performed based on the Random Forest predictor and we achieved the accuracy Ac, Sensitivity Sn, Precession Pr and Matthew Correlation Coefficient MCC for the SARS-COV2 protein data of US-Germany, US-China and China-Germany as presented in the table and the ROC curves are shown in the Table 3 and the ROC curves in Figure 5.

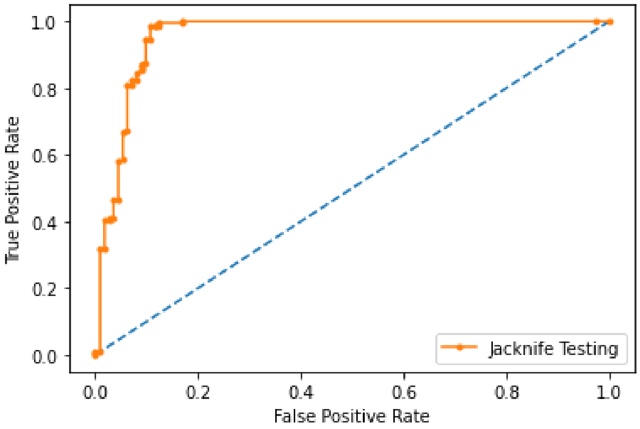
Table 5. Jackknife test result

Model	ACC (%)	MCC	SN (%)	Pr (%)	ROC
US-Germany	95.0	1.0	0.96	1.0	0.95
US-China	96.88	1.0	1.0	1.0	0.96
China-Germany	95.1	1.0	1.0	1.0	0.95



(a).

(b)



(c)

Figure 6. Jackknife testing ROC Curves (a). US-Germany (b). US-China (c). China-Germany

2. Conclusion: In this work, we developed three models to identify protein mutant strains associated with the USA, Germany and China. The model for differentiating the US SARS-COV2 protein strain from the German strain. Similarly, in the second model, the process is repeated for the United States and China, and finally in the third model for China and Germany. The performance of models is evaluated by the independent tests, 10-fold cross-validation, and by the jackknife test. Using these performance evaluations, we achieved very dominant results with an accuracy of around 95% in all models. These results of high accuracy predicted the high level of discrimination in the SARS-COV2 protein strain associated with the three different countries: United States, Germany, and China. This discrimination shows the very visible mutation in Germany and the USA as compared to China, the origin of the virus. Likewise, the visible mutation between the USA and Germany. This concludes the presence of powerful mutation existence in SARS-COV2 protein, which is an alarming situation not only to adopt the controlling strategy but also in case of its drug and vaccine development and its success from area to area. Further, a comprehensive research is required to separately analyses Accessory protein 7b, 6, ORF3a, 10, 8 protein, Envelope small membrane protein, Nucleoprotein, Membrane protein, Spike glycoprotein, 3C-like proteinase, and 2'-O-methyltransferase with respect to different countries so that to identify the most mutated subset of SARS-COV2 protein strain.

REFERENCES

[1]. Tang X., Wu C., Li X., Song Y., Yao X., Wu X., Duan Y., Zhang H., Wang Y., Qian Z. On the origin and continuing evolution of SARS-CoV-2. *Natl. Sci. Rev.* 2020;7(6):1012–1023. [Google Scholar] [Ref list]

[2]. Zhang Y.-Z., Holmes E.C. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell.* 2020;181(2):223–227. [PMC free article] [PubMed] [Google Scholar] [Ref list]

[3]. Zhu, N. *et al.* A novel coronavirus from patients with pneumonia in China, 2019. *N. Eng. J. Med.* 382(8), 727–733 (2020)

[4]. Islam M.R., Hoque M.N., Rahman M.S., Alam A.R.U., Akther M., Puspo J.A., Akter S., Sultana M., Crandall K.A., Hossain M.A. Genome-wide analysis of SARS-CoV-2 virus strains circulating worldwide implicates heterogeneity. Sci. Rep. 2020;10:1–9. [PMC free article] [PubMed] [Google Scholar] [Ref list]

[5]. Li Y., Yang X., Wang N., Wang H., Yin B., Yang X., Jiang W. The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive RNA modification. Futur. Virol. 2020;15(6):341–347.

[6]. Rahman M.S., Islam M.R., Hoque M.N., Alam A.R.U., Akther M., Puspo J.A., Akter S., Anwar A., Sultana M., Hossain M.A. Comprehensive annotations of the mutational spectra of SARS-CoV-2 spike protein: a fast and accurate pipeline. Transbound. Emerg. Dis. 2020:1–13. (2020; 00) [PMC free article] [PubMed] [Google Scholar] [Ref list]

[7]. Walls, A. C. *et al.* Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 180, 1–12 (2020).

[8]. Ahmed, S. F., Quadeer, A. A. & McKay, M. R. Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses* 12(3), 254 (2020).

[9]. WHO Coronavirus disease (COVID-19) Weekly Epidemiological Update and Weekly Operational Update https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports dated April 10, 2021

[10]. Phan, T. Genetic diversity and evolution of SARS-CoV-2. *Infect. Genet. Evol.* 81, 104260 (2020).

[11]. Sardar, R., Satish, D., Birla, S. & Gupta, D. Comparative analyses of SAR-CoV2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. *bioRxiv* (2020).

[12]. Armijos-Jaramillo, V., Yeager, J., Muslin, C. & Perez-Castillo, Y. SARS-CoV-2, an evolutionary perspective of interaction with human ACE2 reveals undiscovered amino acids necessary for complex stability. *bioRxiv* (2020).

[13]. Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, Carter CD, Wheeler D, Davis RW, Boeke JD, Snyder MA, Basrai MA. Functional genomics of genes with small open reading frames (sORFs) in S. cerevisiae. *Genome Res.* 2006;16(3):365–373. [PMC free article] [PubMed] [Google Scholar]

[14]. Basrai MA, Hieter P, Boeke JD. Small open reading frames: beautiful needles in the haystack. *Genome Res.* 1997;7(8):768–771. [PubMed] [Google Scholar]

[15]. Sharp PM, Li WH. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 1987;15(3):1281–1295. [PMC free article] [PubMed] [Google Scholar]

[16]. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS. Global analysis of protein expression in yeast. *Nature.* 2003;425(6959):737–741. [PubMed] [Google Scholar]

[17]. Roy, S., Martinez, D., Platero, H., Lane, T., & Werner-Washburne, M. (2009). Exploiting amino acid composition for predicting protein-protein interactions. *PloS one*, *4*(11), e7813.

[18]. Wang, Y., Zhang, Q., Sun, M. A., & Guo, D. (2011). High-accuracy prediction of bacterial type III secreted effectors based on position-specific amino acid composition profiles. *Bioinformatics*, *27*(6), 777-784.

[19]. Zhou, X. B., Chen, C., Li, Z. C., & Zou, X. Y. (2007). Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *Journal of theoretical biology*, *248*(3), 546-551.

[20]. Chen, C., Shen, Z. B., & Zou, X. Y. (2012). Dual-layer wavelet SVM for predicting protein structural class via the general form of Chou's pseudo amino acid composition. *Protein and peptide letters*, *19*(4), 422-429.

[21]. Chou, K. C. (2009). Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Current Proteomics*, *6*(4), 262-274.

[22]. Shen, H. B., & Chou, K. C. (2008). PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. *Analytical biochemistry*, *373*(2), 386-388.

[23]. Akmal, M. A., Hussain, W., Rasool, N., Khan, Y. D., Khan, S. A., & Chou, K. C. (2020). Using Chou's 5-steps rule to predict O-linked serine glycosylation sites by blending position relative features and statistical moment. *IEEE/ACM transactions on computational biology and bioinformatics*.

[24]. Shah, A. A., & Khan, Y. D. (2020). Identification of 4-carboxyglutamate residue sites based on position based statistical feature and multiple classification. *Scientific Reports*, *10*(1), 1-10.

[25]. Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., & Chou, K. C. (2019). SPalmitoylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-palmitoylation sites in proteins. *Analytical biochemistry*, *568*, 14-23.

[26]. Khan, Y. D., Rasool, N., Hussain, W., Khan, S. A., & Chou, K. C. (2018). iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. *Analytical biochemistry*, *550*, 109-116.

[27]. Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., & Chou, K. C. (2019). SPrenylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins. *Journal of theoretical biology*, *468*, 1-11.

[28]. Awais, M., Hussain, W., Khan, Y. D., Rasool, N., Khan, S. A., & Chou, K. C. (2019). iPhosH-PseAAC: Identify phosphohistidine sites in proteins by blending statistical moments and position relative features according to the Chou's 5-step rule and general pseudo amino acid composition. *IEEE/ACM transactions on computational biology and bioinformatics*.

[29]. Khan, Yaser Daanial, Mehreen Jamil, Waqar Hussain, Nouman Rasool, Sher Afzal Khan, and Kuo-Chen Chou. "pSSbond-PseAAC: Prediction of disulfide bonding sites by integration of PseAAC and statistical moments." *Journal of theoretical biology* 463 (2019): 47-55.

[30]. Khan, S., Khan, M., Iqbal, N., Hussain, T., Khan, S. A., & Chou, K. C. (2019). A two-level computation model based on deep learning algorithm for identification of piRNA and their functions via Chou's 5-steps rule. *International Journal of Peptide Research and Therapeutics*, 1-15.

[31]. A. H. Butt, N. Rasool, and Y. D. Khan, "A Treatise to Computational Approaches Towards Prediction of Membrane Protein and Its Subtypes," *J. Membr. Biol.*, vol. 250, no. 1, pp. 55–76, Feb. 2017, doi: 10.1007/s00232-016-9937-7.

[32]. A. H. Butt, N. Rasool, and Y. D. Khan, "Predicting membrane proteins and their types by extracting various sequence features into Chou's general PseAAC," *Mol. Biol. Rep.*, vol. 45, no. 6, pp. 2295–2306, Dec. 2018, doi: 10.1007/s11033-018-4391-5.

[33]. A. H. Butt, N. Rasool, and Y. D. Khan, "Prediction of antioxidant proteins by incorporating statistical moments based features into Chou's PseAAC," *J. Theor. Biol.*, vol. 473, pp. 1–8, Jul. 2019, doi: 10.1016/j.jtbi.2019.04.019.

[34]. Q. Dai, S. Ma, Y. Hai, Y. Yao, and X. Liu, "A segmentation based model for subcellular location prediction of apoptosis protein," *Chemom. Intell. Lab. Syst.*, vol. 158, pp. 146–154, Nov. 2016, doi: 10.1016/j.chemolab.2016.09.005.

[35]. M. K. & M. Hayat, "iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples," *Mol Genet Genomics*, vol. 291, pp. 285–296 (2016)., 2016, doi: 10.1007/s00438-015-1108-5.

[36]. FarmanAli, MaqsoodHayat, "Classification of membrane protein types using Voting Feature Interval in combination with Chou's Pseudo Amino Acid Composition," *J. Theor. Biol.*, vol. 384, no. 7, pp. 78–83, 2015, doi: 10.1016/j.jtbi.2015.07.034.

[37]. Abhishek Sharma, "Decision Tree vs. Random Forest – Which Algorithm Should you Use?" Retrived: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/ June September 2020.

[38]. Chauhan, A., Chauhan, D., & Rout, C. (2014). Role of gist and PHOG features in computer-aided diagnosis of tuberculosis without segmentation. *PloS one*, *9*(11), e112980.

[39]. Ren L. L., Wang Y. M., Wu Z. Q., Xiang Z. C., Guo L., Xu T., et al. (2020). Identification of a novel coronavirus causing severe pneumonia in human: a descriptive study. *Chin. Med. J.* 133 1015–1024. 10.1097/CM9.0000000000000722

[40]. Choudhry H, Bakhrebah MA, Abdulaal WH, Zamzami MA, Baothman OA, Hassan MA, Zeyadi M, Helmi N, Alzahrani F, Ali A, Zakaria MK, Kamal MA, Warsi MK, Ahmed F, Rasool M, Jamal MS Future Virol. 2019 Apr; 14(4):237-246.

[41]. Brian D. A., Baric R. S. (2005). Coronavirus genome structure and replication. *Curr. Topics Microbiol. Immunol.* 287, 1–30. doi: 10.1007/3-540-26765-4_1

[42]. Jin, Y., Yang, H., Ji, W., Wu, W., Chen, S., Zhang, W., & Duan, G. (2020). Virology, epidemiology, pathogenesis, and control of COVID-19. *Viruses*, *12*(4), 372.

[43]. Olmos, C., Cepeda, J., & Zenteno, D. (2020). NUEVO CORONAVIRUS (COVID-19) EN POBLACIÓN GENERAL Y PEDIÁTRICA: UNA REVISIÓN EPIDEMIOLÓGICA. CHILE 2020. NOVEL CORONAVIRUS (COVID-19) IN GENERAL AND PEDIATRIC POPULATION: AN EPIDEMIOLOGICAL REVIEW. CHILE 2020. *Neumología Pediátrica*, *15*(2), 293-300.

[44]. Sars Cov 2 Virus Genome, https://centri.onrender.com/sars-cov-2-virus-genome.html Retreived April 10, 2021

[45]. Majchrzykiewicz-Koehorst, J. A., Heikens, E., Trip, H., Hulst, A. G., de Jong, A. L., Viveen, M. C., ... & Paauw, A. (2015). Rapid and generic identification of influenza A and other respiratory viruses with mass spectrometry. *Journal of virological methods*, *213*, 75-83.

[46]. Randhawa, G. S., Soltysiak, M. P., El Roz, H., de Souza, C. P., Hill, K. A., & Kari, L. (2020). Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *Plos one*, *15*(4), e0232391.

[47]. Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., & Wei, G. W. (2020). Characterizing SARS-CoV-2 mutations in the United States. *arXiv preprint arXiv:2007.12692*.

[48]. Elizondo, V., Harkins, G. W., Mabvakure, B., Smidt, S., Zappile, P., Marier, C., ... & Duerr, R. (2021). SARS-CoV-2 genomic characterization and clinical manifestation of the COVID-19 outbreak in Uruguay. *Emerging microbes & infections*, *10*(1), 51-65.

[49]. Azad, G. K. (2021). Identification and molecular characterization of mutations in nucleocapsid phosphoprotein of SARS-CoV-2. *PeerJ*, *9*, e10666.

[50]. Khan, Y. D., & Roomi, M. S. (2020). Promising compounds for treatment of Covid-19. *VAWKUM Trans. Comput. Sci*, *17*(1), 1-8.

[51]. Hassan, S. A. (2016). Comparative Computational Analysis of a Putative Transcriptional Regulator Map_PRSO3010 and its implications in the Pathogenesis of Crohn's and Johne's diseases. *VAWKUM Transactions on Computer Sciences*, *4*(1), 60-77.

[52]. Hassan, S. A., & Tayubi, I. A. (2017). Computational Approaches to Identify a Derivative of Galardin as an Inhibitor of Mycobacterial Peptide Deformylase. *VAWKUM Transactions on Computer Sciences*, *5*(1), 45-55.

[53]. Ullah, F., & Khan, I. (2014). Bnmps: Biomolecular nanomachine protocol stack for human disease diagnoses: A new paradigm. *VAWKUM Transactions on Computer Sciences*, *2*(1), 96-106.

[54]. D. S. Cao, Q. S. Xu, and Y. Z. Liang, "Propy: A tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013, doi: 10.1093/bioinformatics/btt072.

[55]. P. Tripathi and P. N. Pandey, "A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition," *J. Theor. Biol.*, vol. 424, pp. 49–54, 2017, doi: 10.1016/j.jtbi.2017.04.027.

[56]. F. Javed and M. Hayat, "Predicting subcellular localization of multi-label proteins by incorporating the sequence

features into Chou's PseAAC," *Genomics*, no. September, pp. 0–1, 2018, doi: 10.1016/j.ygeno.2018.09.004.

[57]. L. Zhang and L. Kong, "iRSpot-ADPM: Identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components," *J. Theor. Biol.*, vol. 441, pp. 1–8, 2018, doi: 10.1016/j.jtbi.2017.12.025.

[58]. Albugami, N. (2020). Prediction of Saudi Arabia SARS-COV 2 Diversifications in Protein Strain Against China Strain. *VAWKUM Transactions on Computer Sciences*, *8*(1), 64-73.

[59]. Hassan, S. A., Khan, T., & Hashmi, A. (2016). Computational Approach to Design Antagonists of Mycobacterium Tuberculosis Lipoprotein Lprg (RV1411C) Protein. *VAWKUM Transactions on Computer Sciences*, *4*(1), 44-50.

[60]. C. Huang and J. Q. Yuan, "Predicting protein subchloroplast locations with both single and multiple sites via three different modes of Chou's pseudo amino acid compositions," *J. Theor. Biol.*, vol. 335, no. 0022, pp. 205–212, 2013, doi: 10.1016/j.jtbi.2013.06.034.

[61]. K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *J. Theor. Biol.*, vol. 273, no. 1, pp. 236–247, 2011, doi: 10.1016/j.jtbi.2010.12.024.

[62]. K. C. Chou, "Prediction of protein cellular attributes using pseudo-amino acid composition," *Proteins Struct. Funct. Genet.*, vol. 43, no. 3, pp. 246–255, 2001, doi: 10.1002/prot.1035.

[63]. X. Fu, W. Zhu, B. Liao, L. Cai, L. Peng et al., "Improved DNA-Binding protein identification by incorporating evolutionary information into the Chou's PseAAC," *IEEE Access*, vol. 6, pp. 66545–66556, 2018, doi: 10.1109/ACCESS.2018.2876656.

[64]. J. Jia, Z. Liu, X. Xiao, B. Liu, and K. C. Chou, "pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach," *J. Theor. Biol.*, vol. 394, pp. 223–230, 2016, doi: 10.1016/j.jtbi.2016.01.020.

[65]. Y. D. Khan, F. Ahmed, and S. A. Khan, "Situation recognition using image moments and recurrent neural networks," *Neural Comput. Appl.*, vol. 24, no. 7–8, pp. 1519–1529, 2014, doi: 10.1007/s00521-013-1372-4.

[66]. W. Hussain, Y. D. Khan, N. Rasool, S. A. Khan, and K. C. Chou, "SPrenylC-PseAAC: A sequence-based model developed via Chou's 5-steps rule and general PseAAC for identifying S-prenylation sites in proteins," *J. Theor. Biol.*, vol. 468, pp. 1–11, 2019, doi: 10.1016/j.jtbi.2019.02.007.