Classifying COVID-19 Spike Sequences from Geographic Location Using Deep Learning

Sarwan Ali[1], Babatunde Bello[1], and Murray Patterson[1]

Georgia State University, Atlanta GA 30303, USA
{sali85,bbello1}@student.gdu.edu,mpatterson30@gsu.edu

Abstract. With the rapid spread of COVID-19 worldwide, viral genomic data is available in the order of millions of sequences on public databases such as GISAID. This *Big Data* creates a unique opportunity for analysis towards the research of effective vaccine development for current pandemics, and avoiding or mitigating future pandemics. One piece of information that comes with every such viral sequence is the geographical location where it was collected — the patterns found between viral variants and geographic location surely being an important part of this analysis. One major challenge that researchers face is processing such huge, highly dimensional data to get useful insights as quickly as possible. Most of the existing methods face scalability issues when dealing with the magnitude of such data. In this paper, we propose an algorithm that first computes a numerical representation of the spike protein sequence of SARS-CoV-2 using k-mers (substrings) and then uses a deep learning-based model to classify the sequences in terms of geographical location. We show that our proposed model significantly outperforms the baselines. We also show the importance of different amino acids in the spike sequences by computing the information gain corresponding to the true class labels.

Keywords: Sequence Classification · SARS-CoV-2 · COVID-19 · k-mers · Deep Learning · Viral Evolution · Geographic Location

1 Introduction

The adaptability of viruses like SARS-CoV-2, when coupled with a variety of selection pressures from the various ecosystems, host immunities and approaches to pharmaceutical intervention provide an evolutionary environment that leads to the emergence of strains and variants in different geographical locations. While SARS-CoV-2 has spread rather quickly to many parts of the globe since the initial outbreak in Wuhan at the end of 2019 which led to the COVID-19 pandemic [42], it continues to raise global concerns as the virus persistently evolves and accumulates new mutations. Consequently, new variants of SARS-CoV-2 have emerged in different parts of the world: the Alpha variant (B.1.1.17) emerged in the UK, Beta (B.1.351) in South Africa, Gamma in Brazil, Epsilon in California, Iota (B.1.526) in New York, Delta (B.1.167.2) and Kappa (B.1.167.1)

in India, to name a few. All of these variants possess some mutations that confer increased transmissibility and higher binding affinity of their spike protein (see Figure 1) to human host ACE2 receptors [14, 19].

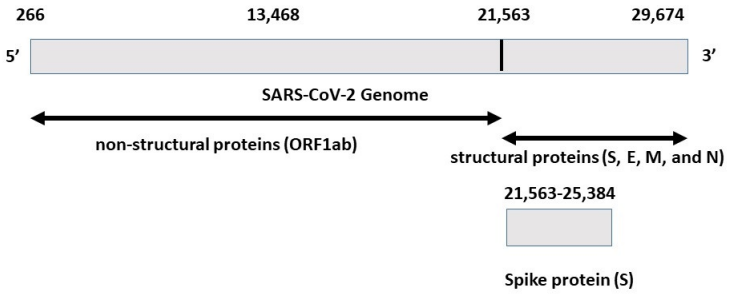

Fig. 1: The SARS-CoV-2 genome codes for several proteins, including the surface, or spike protein. The spike protein is composed of 3821 (25384–21563) nucleotides (and one "stop" character '*'). Therefore, the final length of the spike protein is $3822/3 = 1274$ (we divide by 3 because each amino acid corresponds to 3 DNA characters, or codons) [19].

It is concerning that the longer SARS-CoV-2 has to propagate, its exposure to wider ranges of immune response attacks across diverse communities and geographically diverse environments may be incubating the virus to evolve new variants and strains that are dangerous and extremely immunologically evasive both locally and globally, as the pandemic prolongs. From the point of view of evolution, this is like giving the virus robust evolutionary room and time to learn, to evolve adaptations, gain of function, and escapes from host immune arsenal and attacks. Sadly, this is gradually the case already, as the original Wuhan strain is now almost completely replaced by new variants with different characteristic behaviors and are hence less responsive to the currently available vaccines [18, 20]. This is why it is important to characterize different strains and variants of SARS-CoV-2 based on geographical location, to understand the patterns of spread in hopes to contain, or at least cope with this virus.

The SARS-CoV-2 genome typically accrues 1 or 2 point mutations (SNVs) in a month. According to a review, some 12,706 such mutations have so far been detected by researchers since the advent of the COVID-19 pandemic. While some changes have neutral effects, a few that occur in major proteins are critical to viral evolution, genomic stability, transmissibility, antigenicity, virulence, adaptation and escape from host immune response [27, 29]. The SARS-CoV-2 Spike (S) Protein is a key player in the virus life cycle. The protein is composed of 1274 amino acids encoded by the S-gene of the virus (see Figure 1). It is the major target of the neutralizing antibodies from host immune response and currently available vaccines for COVID-19. The virus uses the spike pro-

tein to bind the host ACE2 receptor on the cell surface (found abundantly in airways, lungs, mucous lines and the intestine) which facilitates the uptake of the virus into host cells [24, 39]. Thus, mutations in the S gene have reportedly imparted viral pathogenesis, binding activity of the spike protein to the host, as well as causing conformational changes in the protein molecule. For instance, mutation D614G was found to enhance the viral infectivity and stability of the SARS-CoV-2 genome, which has been attributed to spike protein assembly on the virion surface [20]. Currently, quite a number of novel variants are being identified by the US Center for Disease Control and Prevention (CDC) and the World Health Organization (WHO) [33]. Since all of these variants are characterized by different spike protein content [14, 19], classification can help us to discover also patterns in the geographic distribution of these variants.

SARS-CoV-2 still circulates among human populations in different locations, weather conditions and epidemiological descriptions. It is important to investigate how this regional diversity contributes to viral evolution and emergence of new variants in these regions. Research suggests possible selective mutations in the SARS-CoV-2 genome: specific sites appear more subject to selective mutation. Some mutational sites in ORF1ab, ORF3a, ORF8 and N region of SARS-CoV-2 reportedly exhibit different rates of mutation [40]. A study involving the analysis and characterization of samples from COVID-19 patients in different parts of the world identify 8 novel recurrent mutational sites in the SARS-CoV-2 genome. Interestingly, the studies also note changes at sites 2891, 3036, 14408, 23403, and 28881 to be common in Europe, while 17746, 17857, and 18060 are common in North America [29]. A recent study also identified the ongoing evolution of SARS-CoV-2 to involve purifying selection, and that a small number of sites appear to be positively selected. The work also identifies the spike protein receptor binding domain (RBD) and a region of nucleocapsid protein to be also positively selected for substitutions. The work also highlighted trend in virus diversity with geographic region and adaptive diversification that may potentially make variant-specific vaccination an issue [32].

Given all of the novel SARS-CoV-2 variants and strains that have emerged from different geographical regions of the world, we need to investigate this connection to the spread of the virus, *e.g.*, weather factors possibly play a systematic role [30, 34]. There is also diversity of immune system across the human population. Genomic variations only cause 20–40% of this immune system variation, while the rest 60–80% is accounted for by age, environment factors like where we live and our neighbors, cohabitation and chronic viral infections, etc. Immune response is also known to show intra-species variation [26]. There is an ongoing evolutionary arms-race between host and pathogens they are exposed to which constantly changes the host anti-pathogen attack and in turn causes the pathogen to refine or adjust its escape from host immune attack [11, 26]. This is constantly taking place, with the virus under evolution pressure and natural selection to propagate the most fit virus. It may be complex to characterize how each factor contribute to this variation. The immune system variation is possibly an important driver on how new variants of SARS-CoV-2 are regionally

emerging with positive selections for escaping immune neutralization, increased infectivity and transmissibility as observed recently.

Classification of the SARS-CoV-2 Spike protein sequences based on geographical location of emergence is therefore an important and informative exploration for possible unique patterns, trends and distribution. SARS-CoV-2 spike protein must interact chemically with the host receptor molecule, ACE2 for cellular uptake. Since millions of spike sequences are available now on public databases like GISAID, classifying those sequences becomes a *Big Data* problem. When dealing with big data, scalability and robustness are two important challenges. Some algorithms are robust while other scale well, but give poor predictive performance on larger datasets. The author of [7] proposed a scalable approach, called Spike2Vec, which is scalable to larger sized dataset. When there is some structure (natural clustering) in the data, Spike2Vec is proven to be useful as compared to one-hot embedding [7]. However, we show in this paper that Spike2Vec does not always work in all types of scenarios. To further improve the results of Spike2Vec and that of one-hot embedding, using Deep learning method was compulsory.

In this paper, we propose to use a simple sequential convolutional neural network along with a k-mers based feature vector representation for classifying the geographical locations of COVID-19 patients using spike protein sequences only. Our contributions in this paper are the following:

1. We show that our deep learning based model is scalable on a high volume of data and significantly outperforms the baseline algorithms.
2. We show the importance of different amino acids within the spike sequence by computing information gain corresponding to the class label.
3. We show that given the complexity of the data, our model is still able to outperform the baselines while using only 1% of the training data.
4. We show that preserving the order of amino acids using k-mers achieve better predictive performance that traditional one-hot encoding based embedding approach.
5. Our approach allows us to predict the geographical region of the COVID infected human while accounting for important local and global variability in the spike sequences.

The rest of the paper is organized as follows: Section 2 contains the related work. The proposed approach is given in Section 3. Dataset detail and experimental setup are in Section 4. The results of our method and comparison with the baseline is shown in Section 5. Finally, we conclude our paper in Section 6.

2 Related Work

Sequence classification is a widely studied problem in domains like sequence homology (shared ancestry) detection between a pair of proteins and Phylogeny based inference [12] of disease transmission [21]. Previous studies on working with fixed length numerical representation of the data successfully perform different data analytics tasks. It has applications in different domains such as

graphs [16, 17], nodes in graphs [6, 15], and electricity consumption [3, 4]. This vector-based representation also achieve significant success in sequence analysis, such as texts [35–37], electroencephalography and electromyography sequences [9, 38], Networks [1], and biological sequences [2]. However, most of the existing sequence classification methods require the input sequences to be aligned. Although sequence alignment help to analyze the data better, it is a very costly process.

Kuzmin et al. in [23] show that viral-host classification can be done efficiently using spike sequences only and applying different machine learning (ML) models. They use one-hot encoding (OHE) to get numerical representation for the spike sequences and then apply traditional ML classifiers after reducing the dimensions of the data using the Principal Component Analysis (PCA) method [41]. Although OHE is proven to be efficient in terms of predictive performance, it does not preserve the order of amino acids in the spike protein if we want to take the pair-wise euclidean distance [5]. Another problem with the one-hot encoding based approach is that it deals with the aligned sequential data.

Many previous studies propose the use of k-mers (substrings of length k), which is an alignment-free approach, instead of traditional OHE based embedding to get the numerical vector representation for the genomic data [5, 7, 8]. After getting substrings of length k, a fixed-length feature vector is generated, containing the count of each unique k-mer in a given sequence. This k-mers based method is used initially for phylogenetic applications [10] and showed success in constructing accurate phylogenetic trees from DNA sequences. Authors in [5] argue that better sequence classification results can be achieved using k-mers instead of OHE because k-mers tends to preserve the order of amino acids within a spike sequence.

After getting the numerical representation, a popular approach is to get the kernel matrix and give that matrix as input to traditional machine learning classifiers like support vector machines (SVM) [13, 22, 25]. Farhan et al. in [13] propose an approximate kernel (Gram matrix) computation algorithm, which uses the k-mers based feature vector representation as an input to kernel computation algorithm.

3 Proposed Approach

In this section, we present our proposed model for classifying regions of people based on spike sequences only. We start by explaining the basic MAJORITY based model for the classification. We then show One-Hot Encoding based feature vector generation approach. After that, we show how we generate k-mers based frequency vectors. In the end, we introduce the deep learning model, which we are using for classification purpose.

3.1 MAJORITY

We start with a simple baseline model called MAJORITY. In this approach, we simply take the class with majority representation in the data and declare it

as the class label for all data points. We then measure the performance of this baseline model using different evaluation metrics.

3.2 One-Hot Encoding [7]

In order to get the numerical representation for the sequence-based data, one of the popular methods is using One-Hot Encoding (OHE) [5, 7, 8, 23]. Note that the length of each spike sequence in our dataset is 1274, which contains characters (amino acids) from a set of 21 unique alphabets "*ACDEFGHIKLMNPQRSTVWXY*". For OHE, since we need to have a 21 dimensional sub-vector for each amino acid, the length of OHE based feature vector for each spike sequence will be $21 \times 1273 = 26,733$ (we take the length of spike protein as 1273 instead of 1274 because we have stopping character '*' at the 1274^{th} position). After getting OHE for the whole data, since the dimensionality of the data will be high, authors in [23] use the typical principal component analysis (PCA) approach for dimensionality reduction. Since the size of data is huge in our case, we simply cannot use PCA because of high computational cost [7]. For this purpose, we use an unsupervised approach for low dimensional feature vector representation, called Random Fourier Features (RFF) [31].

3.3 Random Fourier Features (RFF) Based Embedding

To compute pair-wise similarity between two feature vectors, a popular method is to compute kernel (similarity) matrix (gram matrix) and give it as input to popular classifiers such as support vector machine (SVM) [13]. However, exact kernel methods are expensive in terms of computation (scale poorly on training data [31]), and they require huge space to store an $n \times n$ matrix (where n is the total number of data points). To overcome this problem, we can use the so-called kernel trick.

Definition 1 (Kernel Trick). *It is a fast way to compute the similarity between feature vectors using the inner product. The kernel trick's main goal is to avoid the explicit need to map the input data to a high-dimensional feature space.*

The Kernel Trick relies on the assumption that any positive definite function f(a,b), where $a, b \in \mathcal{R}^d$, defines an inner product and a lifting ϕ so that we can quickly compute the inner product between the lifted data points [31]. It can be described in a formal way using the following expression:

$$\langle \phi(a), \phi(b) \rangle = f(a, b) \tag{1}$$

Although kernel trick is effective in terms of computational complexity, it is still not scalable for multi-million sized data. To overcome these computational and storage problems, we use RFF [31], an unsupervised approach that maps the input data to a randomized low dimensional feature space (euclidean inner product space). It can be described in a formal way using the following expression:

$$z : \mathcal{R}^d \rightarrow \mathcal{R}^D \tag{2}$$

In RFF, we approximate the inner product between a pair of transformed points, which is almost equal to the actual inner product between the original data points. More formally:

$$f(a, b) = \langle \phi(a), \phi(b) \rangle \approx z(a)'z(b) \tag{3}$$

In Equation (3), z is (transformed) low dimensional (approximate) representation of the original feature vector (unlike the lifting ϕ). Since z is the approximate representation of the original feature vector, we can use z as an input for different machine learning (ML) tasks such as classification.

3.4 Spike2Vec

The Spike2Vec is a recently proposed method that uses k-mers and RFF to design low dimensional feature vector representation of the data and then perform typical ML tasks such as classification and clustering [7]. The first step of Spike2Vec is to generate k-mers for the spike sequences.

***k*-mers Computation** The main idea behind k-mers is to preserve the order of amino acids within spike sequences. The k-mers is basically a set of substrings (called mers) of length k. For each spike sequence, the total number of k-mers are the following:

$$\text{Total number of } k\text{-mers} = N - k + 1 \tag{4}$$

where N is the length of spike sequence (1274), and k is a user-defined parameter for the size of each mer. An example of k-mers (where k = 3, 4, and 5) is given in Figure 2. In this paper, we are using $k = 3$ (selected empirically).

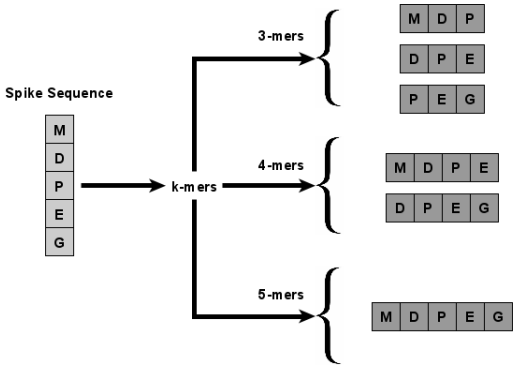

Fig. 2: Example of different length k-mers in a spike sequence "MDPEG".

For OHE and Spike2Vec, we use three classifiers, namely Naive Bayes (NB), Logistic Regression (LR), and Ridge Classifier (RC). For all these classifiers,

default parameters are used for training. To measure the performance, we use average accuracy, precision, recall, weighted and macro F1, receiver operating characteristic curve (ROC) area under the curve (AUC). We also show the training runtime (in sec.) for all methods.

3.5 Keras Classifier

Although Spike2vec is scalable on the multi-million dataset and proved to perform better than typical OHE, it is not efficient in terms of overall predictive performance. To further increase the predictive performance, we use a deep learning-based model called the Keras Classification model (also called Keras classifier). For this purpose, we use a sequential constructor. We create a fully connected network with one hidden layer that contains 9261 neurons (which is equal to the length of the feature vector). The activation function that we are using is "rectifier". In the output layer, we use "softmax" activation function. At the end, we use the efficient Adam gradient descent optimization algorithm with "sparse categorical crossentropy" loss function (used for multi-class classification problem), which computes the crossentropy loss between the labels and predictions. The batch size that we are taking is 100, and we take 10 as number of epochs for training the model. Note that we use OHE and k-mers based frequency vectors (separately) as input for the Keras classifier.

Remark 1. Note that we are using "sparse categorical crossentropy" rather than simple "categorical crossentropy" because we are using integer labels rather than one-hot representation of labels.

4 Experimental Evaluation

4.1 Data Visualization

The spread rate of 3 popular variants of coronavirus (in USA) from March 2020 till July 2021 are given in Figure 3. We can see that the Alpha variant (also known as UK variant [7]) was clearly the variant of concern when it touched its peak in April 2021. We can see the drop in peak for all variants after April 2021. This is because a significant proportion of the population got vaccinated till this point; hence the total cases started decreasing [7]. To evaluate natural clustering in the data (if any exist), we use t-distributed stochastic neighbor embedding [28]. The t-SNE approach maps the data into the 2-dimensional real vector, which can then be visualized using scatter plot. Since applying t-SNE on the whole data is very costly and time consuming, we randomly sampled a subset of data (≈ 80000 sequences) from the data and generated 2D real vector using the t-SNE approach (see Figure 4).

Remark 2. The reason to (randomly) select ≈ 80000 sequences is because t-SNE method is computationally very expensive (runtime is $O(N^2)$, where N is the number of data-points [30]) and take a lot of time on 2.3 million sequences.

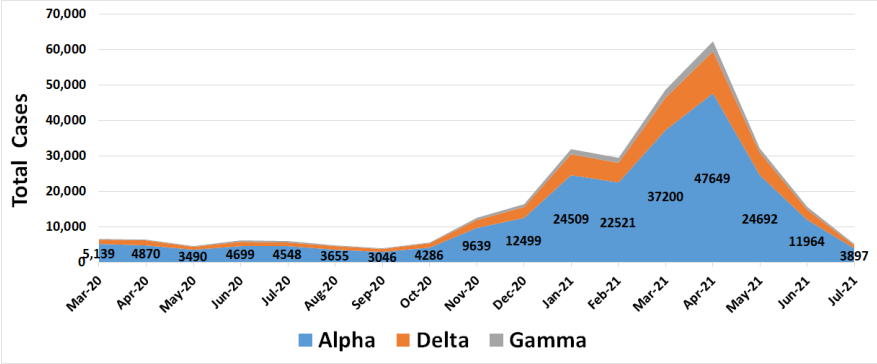

Fig. 3: Spread rate of 3 popular coronavirus variants (in USA) from March 2020 till July 2021.

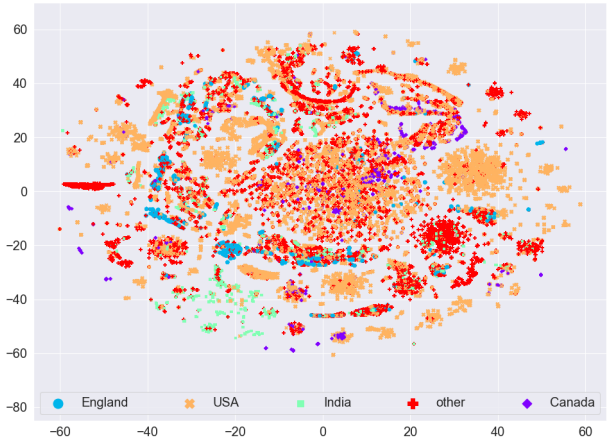

Fig. 4: t-SNE plot for the frequency vectors along with the country information.

4.2 Experimental Setup

All experiments are conducted using an Intel(R) Xeon(R) CPU E7-4850 v4 @ 2.10GHz having Ubuntu 64 bit OS (16.04.7 LTS Xenial Xerus) with 3023 GB memory. Implementation of our algorithm is done in Python and the code is available online for reproducibility[1]. Our pre-processed data is also available online[2], which can be used after agreeing to the terms and conditions of GISAID[3]. For the classification algorithms, we use 1% data for training and 99% for testing. The purpose of using a smaller training dataset is to show how much

[1] Available at https://github.com/sarwanpasha/COVID-19-Country-Classification
[2] Available at https://drive.google.com/drive/folders/1-YmIM8ipFpj-glr9hSF3t6VuofrpgWUa?usp=sharing
[3] Available at https://www.gisaid.org/

performance gain we can achieve while using minimal training data. The dataset statistics are given in Table 1.

Remark 3. Our data split and pre-processing follow those of [5, 7, 8].

Region	Country	Num. of sequences	Region	Country	Num. of sequences
Europe	England	568202	North America	USA	663527
	Germany	146730		Canada	91193
	Denmark	138574		Mexico	20040
	Sweden	78810	Total	3	774760
	Scotland	69387	South America	Brazil	26729
	France	56247	Total	1	26729
	Netherlands	49938	Asia	Japan	75423
	Spain	48830		India	37943
	Switzerland	48516		Israel	14361
	Wales	46851	Total	3	127727
	Italy	44728	Australia	Australia	20985
	Belgium	28758	Total	1	20985
	Ireland	23441			
	Poland	16061			
	Norway	14684			
	Lithuania	13586			
	Luxembourg	12713			
	Finland	11254			
	Slovenia	17135			
Total	19	1434445			

Table 1: The Countries corresponds to 2384646 SARS-CoV-2 spike sequences.

5 Results and Discussion

In this section, we present results for three different granularity of class labels, namely continents, countries, and finally states in a case study of the United States of America (USA).

5.1 Continent Classification

In this section, we show classification results for 5 different continents, namely Europe, North America, South America, Asia, and Australia (see Table 1). The classification results are given in Table 2. In terms of predictive performance, we can observe that the deep learning-based model with k-mers performs best compared to the baselines. While comparing the two embedding methods (i.e.,

OHE and k-mers), we can see that k-mers is better than OHE for the deep learning method and comparable with other ML algorithms. Since k-mers can preserve the order of amino acids more as compared to the OHE, it is able to give richer information in the feature vector. In terms of runtime, RC with k-mers (Spike2Vec [7]) is performing best. The deep learning model will take longer to train the model compared to simple ML classifiers because of the tuning of different parameters.

Approach	Embed. Method	ML Algo.	Acc.	Prec.	Recall	F1 weigh.	F1 Macro	ROC-AUC	Train. runtime (sec.)
MAJORITY	_	_	0.60	0.36	0.60	0.45	0.15	0.50	_
ML Algo.	OHE	NB	0.49	0.63	0.49	0.50	0.38	0.63	1457.2
		LR	0.67	0.66	0.67	0.64	0.33	0.58	1622.4
		RC	0.67	0.66	0.67	0.64	0.28	0.57	1329.1
	Spike2Vec	NB	0.48	0.63	0.48	0.49	0.36	0.63	970.6
		LR	0.67	0.67	0.67	0.64	0.34	0.58	1141.9
		RC	0.67	0.66	0.67	0.64	0.29	0.57	**832.3**
Deep Learning	One-Hot	Keras Classifier	0.75	0.76	0.75	0.72	0.47	0.65	30932.0
	k-mers	Keras Classifier	**0.77**	**0.78**	**0.77**	**0.74**	**0.49**	**0.65**	18631.7

Table 2: Continent Classification Results (1% training set and 99% testing set) for 5 continents (2384646 spike sequences). Best values are shown in bold.

5.2 Country Classification

After classifying the continents, we take countries as the class label and train all ML and deep learning models again with the same parameter settings. The classification results for countries is given in Table 3. In terms of predictive performance, we can observe that the deep learning-based model is performing better than all baselines. In terms of runtime, RC with OHE is the best classifier. An important observation here is the drop in overall performance of all classification models as compared to the continent classification. The reason for this behavior is that there is no such natural clustering or other information in the spike sequences corresponding to the location of patients (see Figure 4). This lack of knowledge in data makes country classification a difficult task. However, we can see that the deep learning-based model can still classify the countries better than the baselines.

5.3 A Case Study of the United States of America (USA)

After classifying continents and countries, we investigate our model with more high granular class labels. For this purpose, we first take the single country with

Approach	Embed. Method	ML Algo.	Acc.	Prec.	Recall	F1 weigh.	F1 Macro	ROC-AUC	Train. runtime (sec.)
MAJORITY	_	_	0.27	0.07	0.27	0.12	0.01	0.5	_
ML Algo.	OHE	NB	0.11	0.44	0.11	0.11	0.10	0.55	1308.4
		LR	0.40	0.46	0.40	0.33	0.15	0.55	2361.8
		RC	0.40	0.38	0.40	0.31	0.11	0.54	**746.4**
	Spike2Vec	NB	0.13	0.41	0.13	0.151	0.109	0.555	1315.3
		LR	0.40	0.45	0.40	0.33	0.16	0.55	2736.8
		RC	0.39	0.37	0.39	0.31	0.11	0.54	779.4
Deep Learning	One-Hot	Keras Classifier	0.49	0.53	0.49	0.43	0.24	0.6	28914.8
	k-mers	Keras Classifier	**0.51**	**0.57**	**0.51**	**0.45**	**0.28**	**0.60**	10383.6

Table 3: Country Classification Results (1% training set and 99% testing set) for 27 countries (2384646 spike sequences). Best values are shown in bold.

the highest spike sequences in the data. Since the USA contains most of the spike sequences in our data (see Table 1), we took it as a case study to further explore different states within the USA. The pie chart showing the distribution of top affected states of the USA are given in Figure 5. The classification results for different states are given in Table 4. We can again observe the drop in predictive performance for all ML and Deep learning models. This again proves that as we increase the granularity of the class labels, it becomes difficult for any model to classify with higher accuracy. We can also observe that the deep learning based model with the k-mers is performing better than all the baselines.

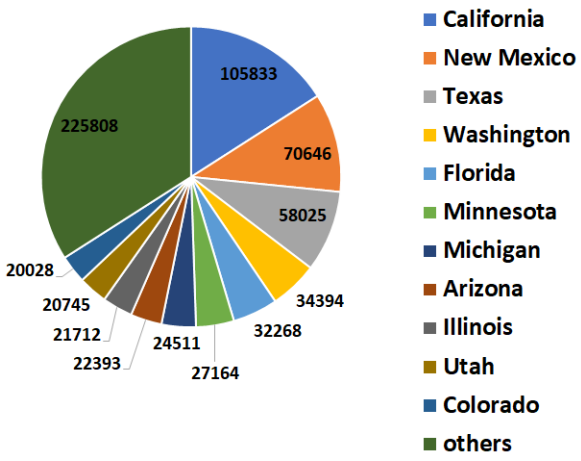

Fig. 5: Distribution of USA's states. Total number of sequences are 663527.

Approach	Embed. Method	ML Algo.	Acc.	Prec.	Recall	F1 weigh.	F1 Macro	ROC-AUC	Train. runtime (sec.)
MAJORITY	_	_	0.33	0.11	0.33	0.17	0.04	0.50	_
ML Algo.	OHE	NB	0.18	0.32	0.18	0.14	0.13	0.54	860.2
		LR	0.37	0.45	0.37	0.26	0.13	0.53	1036.2
		RC	0.37	0.41	0.37	0.25	0.12	0.52	707.7
	Spike2Vec	NB	0.19	0.37	0.19	0.14	0.14	0.55	273.7
		LR	0.38	0.44	0.38	0.29	0.16	0.54	374.2
		RC	0.37	0.42	0.37	0.27	0.14	0.53	**197.1**
Deep Learning	One-Hot	Keras Classifier	0.38	0.44	0.38	0.34	0.22	0.57	7881.3
	k-mers	Keras Classifier	**0.47**	**0.50**	**0.47**	**0.42**	**0.33**	**0.61**	4908.6

Table 4: Classification results for different states of USA (1% training set and 99% testing set). The best values are shown in bold.

5.4 Importance of Attributes

To evaluate importance positions in spike sequences, we find the importance of each attribute with respect to class label (using Weka tool[4]). For this purpose, a randomly selected subset of spike sequences ($\approx 80,000$) is taken from the original dataset. We then compute the Information Gain (IG) between each attribute (amino acid) and the true class label (country). More formally, IG can be computed as follows:

$$IG(Class, position) = H(Class) - H(Class|position) \tag{5}$$

where $H(Class)$ and $H(Class|position)$ are entropy and conditional entropy, respectively. The entropy H can be calculated using following expression:

$$H = \sum_{i \in Class} -p_i \log p_i \tag{6}$$

where p_i is the probability of the class i. The IG values for each attribute is given in Figure 6. The IG values for each attribute is also available online[5].

6 Conclusion

This paper proposes a deep learning-based model that uses k-mers based representation as input and efficiently classifies the COVID-19 patients using spike

[4] Available at https://www.cs.waikato.ac.nz/ml/weka/

[5] Available at https://github.com/sarwanpasha/COVID-19-Country-Classification/blob/main/attributes_correlation.csv

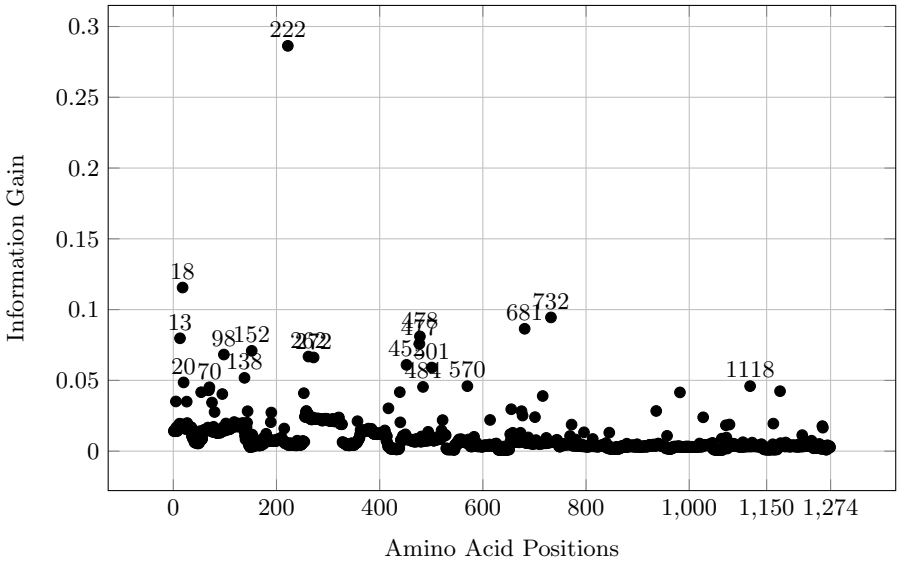

Fig. 6: Information Gain for each amino acid position corresponding to the class.

sequences only. We show that our proposed algorithm is outperforming the baselines in terms of predictive performance. Using the information gain, we also show the importance of attributes (amino acids) in the spike sequences. In the future, we will explore more sophisticated models like LSTM and GRU and also use other attributes like months information to increase the predictive performance. Using other alignment-free methods such as Minimizers is another possible future direction.

References

1. Ali, S., Alvi, M., Faizullah, S., Khan, M., Alshanqiti, A., Khan, I.: Detecting ddos attack on sdn due to vulnerabilities in openflow. In: International Conference on Advances in the Emerging Computing Technologies (AECT). pp. 1–6 (2020)
2. Ali, S., Ciccolella, S., Lucarella, L., D. Vedova, G., Patterson, M.D.: Simpler and faster development of tumor phylogeny pipelines. bioRxiv 458137 (2021)
3. Ali, S., Mansoor, H., Arshad, N., Khan, I.: Short term load forecasting using smart meter data. In: International Conference on Future Energy Systems (e-Energy). pp. 419–421 (2019)
4. Ali, S., Mansoor, H., Khan, I., Arshad, N., Khan, M., Faizullah, S.: Short-term load forecasting using ami data. CoRR **abs/1912.12479** (2020)
5. Ali, S., Sahoo, B., Ullah, N., Zelikovskiy, A., Patterson, M.D., Khan, I.: A k-mer based approach for sars-cov-2 variant identification. Accepted at International Symposium on Bioinformatics Research and Applications (ISBRA) (2021)
6. Ali, S., Shakeel, M., Khan, I., Faizullah, S., Khan, M.: Predicting attributes of nodes using network structure. ACM Transactions on Intelligent Systems and Technology (TIST) **12**(2), 1–23 (2021)

7. Ali, S., Patterson, M.: Spike2vec: An efficient and scalable embedding approach for covid-19 spike sequences. arXiv preprint arXiv:2109.05019 (2021)
8. Ali, S., Tamkanat-E-Ali, Khan, M.A., Khan, I., Patterson, M., et al.: Effective and scalable clustering of sars-cov-2 sequences. Accepted at International Conference on Big Data Research (ICBDR) (2021)
9. Atzori, M., et al.: Electromyography data for non-invasive naturally-controlled robotic hand prostheses. Sci. data **1**(1), 1–13 (2014)
10. Blaisdell, B.: A measure of the similarity of sets of sequences not requiring sequence alignment. Proceedings of the National Academy of Sciences **83**, 5155–5159 (1986)
11. Brodin, P., Jojic, V., Gao, T., Bhattacharya, S., Angel, C.J.L., Furman, D., Shen-Orr, S., Dekker, C.L., Swan, G.E., Butte, A.J., et al.: Variation in the human immune system is largely driven by non-heritable influences. Cell **160**(1-2), 37–47 (2015)
12. Dhar, S., et al.: Tnet: Phylogeny-based inference of disease transmission networks using within-host strain diversity. In: International Symposium on Bioinformatics Research and Applications (ISBRA). pp. 203–216 (2020)
13. Farhan, M., Tariq, J., Zaman, A., Shabbir, M., Khan, I.: Efficient approximation algorithms for strings kernel based sequence classification. In: Advances in neural information processing systems (NeurIPS). pp. 6935–6945. . (2017)
14. Farinholt, T., Doddapaneni, H., Qin, X., Menon, V., Meng, Q., Metcalf, G., Chao, H., Gingras, M.C., Farinholt, P., Agrawal, C., et al.: Transmission event of sars-cov-2 delta variant reveals multiple vaccine breakthrough infections. medRxiv (2021)
15. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: International Conference on Knowledge Discovery & Data Mining (KDD). pp. 855–864 (2016)
16. Hassan, Z., Khan, I., Shabbir, M., Abbas, W.: Computing graph descriptors on edge streams (2021), https://www.researchgate.net/publication/353671195_Computing_Graph_Descriptors_on_Edge_Streams
17. Hassan, Z., Shabbir, M., Khan, I., Abbas, W.: Estimating descriptors for large graphs. In: Advances in Knowledge Discovery and Data Mining (PAKDD). pp. 779–791 (2020)
18. Hu, B., Guo, H., Zhou, P., Shi, Z.L.: Characteristics of sars-cov-2 and covid-19. Nature Reviews Microbiology **19**(3), 141–154 (2021)
19. Huang, Y., Yang, C., Xu, X.f., Xu, W., Liu, S.w.: Structural and functional properties of sars-cov-2 spike protein: potential antivirus drug development for covid-19. Acta Pharmacologica Sinica **41**(9), 1141–1149 (2020)
20. Korber, B., Fischer, W.M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E.E., Bhattacharya, T., Foley, B., et al.: Tracking changes in sars-cov-2 spike: evidence that d614g increases infectivity of the covid-19 virus. Cell **182**(4), 812–827 (2020)
21. Krishnan, G., Kamath, S., Sugumaran, V.: Predicting vaccine hesitancy and vaccine sentiment using topic modeling and evolutionary optimization. In: International Conference on Applications of Natural Language to Information Systems (NLDB). pp. 255–263 (2021)
22. Kuksa, P., Khan, I., Pavlovic, V.: Generalized similarity kernels for efficient sequence classification. In: SIAM International Conference on Data Mining (SDM). pp. 873–882 (2012)
23. Kuzmin, K., et al.: Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone. Biochemical and Biophysical Research Communications **533**, 553–558 (2020)

24. Lamers, M.M., Beumer, J., van der Vaart, J., Knoops, K., Puschhof, J., Breugem, T.I., Ravelli, R.B., Van Schayck, J.P., Mykytyn, A.Z., Duimel, H.Q., et al.: Sars-cov-2 productively infects human gut enterocytes. Science **369**(6499), 50–54 (2020)
25. Leslie, C., Eskin, E., Weston, J., Noble, W.: Mismatch string kernels for svm protein classification. In: Advances in neural information processing systems (NeurIPS). pp. 1441–1448 (2003)
26. Liston, A., Carr, E.J., Linterman, M.A.: Shaping variation in the human immune system. Trends in immunology **37**(10), 637–646 (2016)
27. Lorenzo-Redondo, R., Nam, H.H., Roberts, S.C., Simons, L.M., Jennings, L.J., Qi, C., Achenbach, C.J., Hauser, A.R., Ison, M.G., Hultquist, J.F., et al.: A unique clade of sars-cov-2 viruses is associated with lower viral loads in patient upper airways. MedRxiv (2020)
28. Van der M., L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research (JMLR) **9**(11) (2008)
29. Pachetti, M., Marini, B., Benedetti, F., Giudici, F., Mauro, E., Storici, P., Masciovecchio, C., Angeletti, S., Ciccozzi, M., Gallo, R.C., et al.: Emerging sars-cov-2 mutation hot spots include a novel rna-dependent-rna polymerase variant. Journal of translational medicine **18**(1), 1–9 (2020)
30. Pezzotti, N., Lelieveldt, B.P., Van Der Maaten, L., Höllt, T., Eisemann, E., Vilanova, A.: Approximated and user steerable tsne for progressive visual analytics. IEEE transactions on visualization and computer graphics **23**(7), 1739–1752 (2016)
31. Rahimi, A., Recht, B., et al.: Random features for large-scale kernel machines. In: NIPS. p. 5 (2007)
32. Rochman, N.D., Wolf, Y.I., Faure, G., Mutz, P., Zhang, F., Koonin, E.V.: Ongoing global and regional adaptive evolution of sars-cov-2. Proceedings of the National Academy of Sciences **118**(29) (2021)
33. SARS-CoV-2 Variant Classifications and Definitions: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html (2021), [Online; accessed 1-October-2021]
34. Segovia-Dominguez, I., Zhen, Z., Wagh, R., Lee, H., Gel, Y.R.: Tlife-lstm: Forecasting future covid-19 progression with topological signatures of atmospheric conditions. In: Advances in Knowledge Discovery and Data Mining. pp. 201–212. Springer International Publishing, Cham (2021)
35. Shakeel., M., Karim, A., Khan, I.: A multi-cascaded deep model for bilingual sms classification. In: International Conference on Neural Information Processing (ICONIP). pp. 287–298 (2019)
36. Shakeel, M., Karim, A., Khan, I.: A multi-cascaded model with data augmentation for enhanced paraphrase detection in short texts. Information Processing & Management **57**, 1–19 (2020)
37. Shakeel, M.H., Faizullah, S., Alghamidi, T., Khan, I.: Language independent sentiment analysis. In: International Conference on Advances in the Emerging Computing Technologies (AECT). pp. 1–5 (2020)
38. Ullah, A., Ali, S., Khan, I., Khan, M., Faizullah, S.: Effect of analysis window and feature selection on classification of hand movements using emg signal. In: SAI Intelligent Systems Conference (IntelliSys). pp. 400–415 (2020)
39. V'kovski, P., Kratzel, A., Steiner, S., Stalder, H., Thiel, V.: Coronavirus biology and replication: implications for sars-cov-2. Nature Reviews Microbiology **19**(3), 155–170 (2021)
40. Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., Zhang, Z.: The establishment of reference sequence for sars-cov-2 and variation analysis. Journal of medical virology **92**(6), 667–674 (2020)

41. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)
42. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., et al.: A new coronavirus associated with human respiratory disease in china. Nature **579**, 265–269 (2020)