Spike2Vec: An Efficient and Scalable Embedding Approach for COVID-19 Spike Sequences

1st Sarwan Ali
Department of Computer Science
Georgia State University
Atlanta, GA, USA
sali85@student.gsu.edu

2nd Murray Patterson
Department of Computer Science
Georgia State University
Atlanta, GA, USA
mpatterson30@gsu.edu

***Abstract*—With the rapid global spread of COVID-19, more and more data related to this virus is becoming available, including genomic sequence data. The total number of genomic sequences that are publicly available on platforms such as GISAID is currently several million, and is increasing with every day. The availability of such *Big Data* creates a new opportunity for researchers to study this virus in detail. This is particularly important with all of the dynamics of the COVID-19 variants which emerge and circulate. This rich data source will give us insights on the best ways to perform genomic surveillance for this and future pandemic threats, with the ultimate goal of mitigating or eliminating such threats. Analyzing and processing the several million genomic sequences is a challenging task. Although traditional methods for sequence classification are proven to be effective, they are not designed to deal with these specific types of genomic sequences. Moreover, most of the existing methods also face the issue of scalability. Previous studies which were tailored to coronavirus genomic data proposed to use spike sequences (corresponding to a subsequence of the genome), rather than using the complete genomic sequence, to perform different machine learning (ML) tasks such as classification and clustering. However, those methods suffer from scalability issues.**

In this paper, we propose an approach called Spike2Vec, an efficient and scalable feature vector representation for each spike sequence that can be used for downstream ML tasks. Through experiments, we show that Spike2Vec is not only scalable on several million spike sequences, but also outperforms the baseline models in terms of prediction accuracy, F1-score, etc. Since this type of study on such huge numbers of spike sequences has not been done before (to the best of our knowledge), we believe that it will open new doors for researchers to use this data and perform different tasks to unfold new information that was not available before. As an example of this, we use information gain (IG) to compute the importance of each amino acid in the spike sequence. The amino acids with higher IG values tend to be the same to the ones reported by the USA based Centers for Disease Control and Prevention (CDC) for different variants.

***Index Terms*—COVID-19 Spike Sequences, Feature Vector Representation, k-mers, Classification, Clustering**

I. Introduction

Very few fields of study remain untouched in the big data era, as massive amounts of data are collected in every domain from finance [1], [2] to astronomy [3], [4]. The field of biomedical and health informatics is no exception; one which has had a recent and rather rapid growth spurt in the amount of available data, due to the COVID-19 pandemic [5], [6]. One facet of this increase is the amount of genomic data becoming available for COVID-19 in databases such as GISAID[1], where several million viral genome (virome) sequences of COVID-19 — or more precisely, SARS-CoV-2[2] — are available.

Such data has a high *volume*, as the SARS-CoV-2 virome has $\approx$ 30K nucleotide base-pairs, and there are more than 2.5 million such sequences available in GISAID alone. While the number COVID-19 patients being sequenced is a fraction of the actual number of cases, the sheer number of infections (both now and in the past) means that the *velocity* in which SARS-CoV-2 virome sequences are appearing is very high. For example, in March 2020, when COVID-19 was declared a pandemic by the world health organization (WHO), there were a few thousand sequences available. This grew to tens of thousands in the late summer/fall, when the Alpha variant emerged in the UK [7]. By the end of 2020 it was hundreds of thousands, and in early 2021 it had reached 1 million; today it is over 2.5 million. This will likely continue to increase exponentially (see [8]) as many countries [9], [10] ramp up their sequencing infrastructure for COVID-19 and future pandemics.

The available SARS-CoV-2 virome sequence data, in databases such as GISAID, has a high *variety*, since it comprises sequences from all over the world. Since COVID-19 has spread all over the world for more than a year now, and viruses continue to mutate over time, there are quite a number of variants of the SARS-CoV-2 virome, and they continue to emerge. Because variants continue to emerge and die off, some epidemiologists have even proposed a dynamic nomenclature system similar to that used for the common cold or flu [11]. We use this so-called "Pango Lineage" nomenclature to identify the variants we study here, since only the very common variants of concern (VoCs) are named. Examples of such named VoCs are the Alpha [7] (Pango Lineage B.1.1.7), Gamma [12] (P.1) and Delta [13] (B.1.617.2) variants (see Table I for a more complete list). The genomic variations (most of which happen in the spike region, see Figure 1) that define these different variants have been associated with increased transmissibility [14], and immune evasion [15].

[1]www.gisaid.org

[2]Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), the virus which causes the COVID-19 disease

Because databases such as GISAID collect sequences from all over the world, they come from heterogeneous sources of sequencing technologies and centers, leading to multiple levels of *veracity*. However, the largest source of different veracity in the data is the widely varying degree to which different populations are represented. For example, the UK sequences about 5% of its population of $\approx$ 70 million, the USA sequences about 1% of its population of $\approx$ 300 million, while India sequences only a fraction of a percent of its population of $\approx$ 1.3 billion [16], [17] (see Figure 3). Because of this, for example, even though the Delta variant likely originated in India, the majority of the available sequences of this variant are from the UK and the USA, after it arrived in these countries [18].

Since the genomic sequence of a virus encodes all of its functions such as virulence and transmissibility, the *value* of such massive amounts of genomic data is clear. It is variation in this genomic sequence itself which defines the different variants of SARS-CoV-2 such as Alpha, Delta and Gamma. All of these variants differ from each other in effect (due to their unique genomic variations), yet they all descend from the original SARS-CoV-2 sequence [19]. It is only through a process of evolution and transmission to many parts of the globe for over a year, has it diverged to this extent. The amount of sequence data available today puts us in the age of *genomic surveillance*: tracking the spread of pathogens in terms of genomic content [20], [21].

Approaches for rapidly clustering and classifying sequences will be crucial in these genomic surveillance efforts. A clustering method, when applied to the data on a daily basis, for example, would identify a new and rapidly emerging variant in terms of a cluster which grows abnormally quickly, allowing scientists to focus on this cluster. Classification, on the other hand, would allow us to track the spread of known variants in new municipalities, regions, countries and continents. For example, the USA had a wave of the Alpha variant from the UK in early 2021, and later, a wave of the Delta variant from India and/or via other intermediaries, such as the UK (see Figure 5). Such patterns of spread can reveal information about the underlying transmission networks between different countries (the UK and USA, or the UK and India), or even parts of different countries. This can help overcome some of the different veracity in the data, such as the widely varying degree to which different countries are represented in terms of sequencing data, due to sampling bias. For example, even though India is very under-sampled compared to the UK, the wave of the Delta variant in the UK, along with information about flights from India to the UK near the beginning of this wave could give us insights on how the Delta variant originated and spread in India.

The development of clustering and classification approaches needs several important considerations, however. For one, the number of sequences is so huge that any way of extracting useful features becomes even more critical. Since the spike protein is the entry point of the virus to the host cell, it is an important characterizing feature of a coronavirus [22], [23]. Most of the variants of SARS-CoV-2 are characterized by mutations which happen disproportionately in the spike region of the genome [7], [12], [13]. Even the mRNA vaccines (*e.g.*, Pfizer and Moderna) for COVID-19 are designed to encode/target only the SARS-CoV-2 spike protein [24] (unlike traditional vaccines which comprise an entire virome). Since the spike region is sufficient to characterize most of the important features of a viral sample, yet much smaller in length, as depicted in Figure 1, we focus on an embedding approach tailored to the spike region of the sequences.

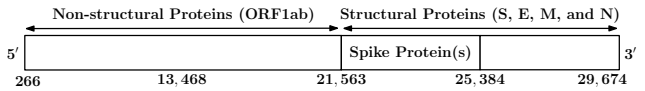

Fig. 1: The SARS-CoV-2 genome is composed of $\approx$ 30Kb nucleotide base pairs, which codes for several proteins, including the spike protein. The region of the genome which corresponds to the spike protein is composed of 3821 $(25,384 - 21,563)$ nucleotide base pairs, hence 3821 (+ 1 stop codon *) $/3 =$ 1274 amino acids.

Previously, some efforts has been done to perform classification and clustering of SARS-CoV-2 spike sequences [25]–[27]. However, those methods are not scalable to the amount of data we use in this study. Although they were successful in getting higher predictive accuracy, it is not clear if the proposed methods are robust and will give same predictive performance on larger datasets. In this paper, we propose Spike2Vec, an efficient and scalable feature vector generation approach for SARS-CoV-2 spike sequences, to which we can apply different machine learning tasks downstream, such as classification and clustering. Our contributions in this paper are as follows:

1) We propose an embedding approach, called Spike2Vec that outperforms the baseline classification method in terms of predictive accuracy.
2) We show that our method is scalable on larger datasets by using ≈ 2.5 million spike sequences.
3) We prove from the results that the machine learning models used in [25]–[27] are not scalable on these larger datasets. This robust checking help us to analyze the machine learning models in detail in terms of their appropriateness for SARS-CoV-2 spike sequences.
4) We also show that in terms of clustering, our embedding approach is better than the baseline model.

The rest of the paper is organized as follows: Section II contains a discussion on the previous studies related to our research problem. Section III contains a detailed description of our Spike2Vec approach. Section IV contains the implementation details of the experimental evaluation of Spike2Vec, along with the dataset statistics and discussion of the baseline models. We present and discuss the results of this experimental evaluation in Section V. Finally, we conclude our paper in Section VI.

II. Literature Review

Because of the rapid spread of COVID-19 since December 2019, a lot of sequence data is available for this virus. This new source of information has attracted researchers from all the fields to perform analysis on this data to better understand the diversity and dynamics of this virus. Authors in [27] propose a one-hot encoding based approach to classify the different coronavirus hosts using spike sequences alone. They shows that near-optimal prediction accuracy can be achieved by considering only the spike portion of the genome sequence rather than using the whole sequence. Ali *et al.*, in [25] perform classification of different variants of the human SARS-CoV-2. Although they were successful in achieving higher accuracy than in [27], the kernel method used in their approach, however, is not scalable to the size of the data we use in this study. This drawback makes it difficult to use this approach in real-world scenario, *i.e.*, the current scenario.

Supervised and unsupervised feature selection methods such as ridge regression [28], lasso regression, and principal component analysis (PCA) [29], etc., are very popular for not only reducing the runtime but also for improving the predictive performance of the underlying machine learning algorithms. Authors in [26] performs clustering on SARS-CoV-2 spike sequences and show that clustering performance could be improved by simply using lasso and ridge regression. Although they were also able to get significant improvement in terms of clustering quality as compared to the baseline, using feature selection methods like ridge regression and lasso regression scales very poorly on the larger datasets, such as the one we use in this study. Melnyk *et al.*, in [30] perform clustering of the entire SARS-CoV-2 genome (rather than just the spike sequence) using CliqueSNV [31], a method originally designed for identifying haplotypes in an intra-host viral population. Although they obtained good overall F1-scores, our (clustering) approach tends to obtain better overall F1-scores. It would be interesting to know whether that is because of our feature vector representation, or because we leverage more (and more up-to-date) data, or both.

Farhan *et al.*, in [32] propose an efficient approach to compute a similarity matrix (kernel matrix). The computed kernel matrix is proven to be efficient for sequence classification. However, since their approach requires to save an entire $n \times n$ dimensional kernel matrix (where n is the total number of sequences), this makes their method expensive in terms of space. Authors in [33] use the random feature method to map the original input into low dimensional feature space so that the inner product of the low dimensional data is approximately equals to the inner product of original data points. Peng *et al.*, in [34] use the random feature attention model for text classification. Their approach is linear in terms of runtime and space and uses random feature methods to approximate the softmax function.

While dealing with *Big Data*, it is also important to analyze the trade-off between the prediction accuracy and the runtime [35]. Although Ali *et al.*, in [25] use the kernel method for spike sequence classification, since the kernel computation is, however, expensive in terms of time and space, their approach is only a proof of concept, and not feasible in a real-world scenario.

III. Proposed Approach

In this section, we give a step by step description of the Spike2Vec approach. Given the SARS-CoV-2 spike sequences, we first generate k-mers so that we can preserve some ordering information of the sequences. An interesting side note is that *de novo* genome assembly (when no reference is present) involves inferring sequence order by assembling k-mers obtained from short reads [36]. After the k-mers are generated, to convert the alphabetical information of k-mers into numerical representation (so that ML algorithms can be applied), we generate frequency vectors, which count the number of occurrences of each k-mer in the spike sequence. We then map the high dimensional frequency vectors to low dimensional embedding using an approximate kernel approach. Each step of Spike2Vec is explained in detail below.

A. k-mers Generation

The first step of Spike2Vec is to compute all k-mers for the spike sequences. The main idea behind using k-mers is to allow some order information of the sequence to be preserved. The total number of k-mers, which we can be generated from a spike sequence of length N is:

$$N - k + 1 \tag{1}$$

In our data, the value for N is 1274. In Equation (1), k is a user-defined parameter for the size of each mer. See Figure 2 for an example. In our experiments we use $k = 3$; this was decided using a standard validation set approach [37].

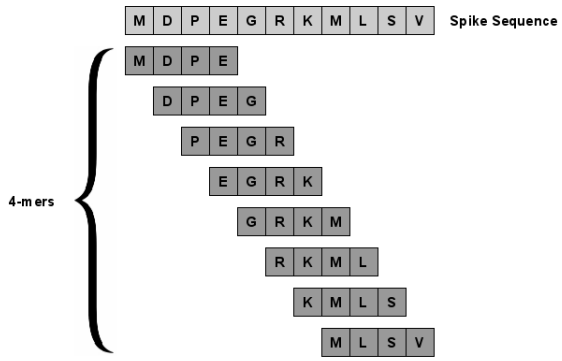

Fig. 2: Example of 4-mers of the amino acid sequence "MDPEGRKMLSV".

B. Frequency Vectors Generation

Since a k-mers is an alphabets-based representation of a spike sequences, we need to convert the k-mers into a numerical representation. Therefore, we design feature vectors that contain the frequency/counts of each k-mer in its respective spike sequence. Each sequence A is over an alphabet Σ.

Remark 1: Alphabets in our dataset represents amino acids of the spike sequence.
These fixed length frequency vector have length $|\Sigma|^k$, which represents the number of possible k-mers of a spike sequence. Since the total number of alphabets in our data are 21 (the number of amino acids), the length of each frequency vector becomes $21^3 = 9261$.

C. Low dimensional Representation

In large scale machine learning (ML) tasks such as classification and regression, typical supervised/unsupervised dimensionality reduction methods such as principal component analysis, ridge regression, and lasso regression, etc., are not suitable because they take a lot more time to execute. Therefore, in a real world scenario where we can have a huge amount of data, the scalability of any underlying algorithm could be one of the major issues. One option is to use kernel based algorithms that compute a similarity matrix which can later be used for the underlying ML tasks. To compute the kernel matrix (gram matrix), the kernel trick is used.

Definition 3.1 (Kernel Trick): The Kernel Trick is used to generate features for an algorithm which depends on the inner product between only the pairs of input data points. The main idea is to avoid the need to map the input data (explicitly) to a high-dimensional feature space.
The Kernel Trick relies on the following observation: *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* [33]. More formally:

$$\langle \phi(a), \phi(b) \rangle = f(a, b) \tag{2}$$

The major drawback of kernel methods is that in case of large training data, the kernel method suffers from large initial computational and storage costs.

To overcome these computational and storage problems, we use an approximate kernel method called Random Fourier Features (RFF) [33], which maps the input data to a randomized low dimensional feature space (euclidean inner product space). More formally:

$$z : \mathcal{R}^d \rightarrow \mathcal{R}^D \tag{3}$$

In this way, we approximate the inner product between a pair of transformed points. More formally:

$$f(a, b) = \langle \phi(a), \phi(b) \rangle \approx z(a)' z(b) \tag{4}$$

In Equation (4), z is low dimensional (unlike the lifting ϕ). In this way, we can transform the original input data with z, which acts as the approximate low dimensional embedding for the original data. This low dimensional representation is then used as an input for different ML tasks like classification and regression.

IV. Experimental Evaluation

We now detail the experiments we performed to evaluate Spike2Vec in terms of both the downstream classification and clustering results obtained.

A. 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 Spike2Vec is done in Python and the code is available online for reproducibility[3]. Our pre-processed data is also available online[4], which can be used after agreeing to terms and conditions of GISAID[5]. For the classification algorithms, we use 1% data for training and 99% for testing. The purpose of using smaller training dataset is to show how much performance gain we can achieve while using minimal training data.

Remark 2: Our data split and pre-processing follow those of [25].

B. Dataset Statistics

We used the (aligned) amino acid sequences corresponding to the spike protein from the largest known database of SARS-CoV-2 sequences, GISAID. In our dataset, we have 2,519,386 spike sequences along with the COVID-19 variant information (in our data, we have 1327 variants in total) for each spike sequence. The information about some of the more well-represented variants is given in Table I. Since most of the variants are new, we do not have all the information available for all them. Therefore, we put "-" in the field in Table I for which we do not have any information available online.

Figure 3 shows the total number of spike sequences for the top 10 countries worldwide. In our GISAID dataset, a total of 219 countries are represented. Since USA has the highest number of spike sequences, we use it as a case study to analyze the spread patterns of different variants in Section IV-C1.

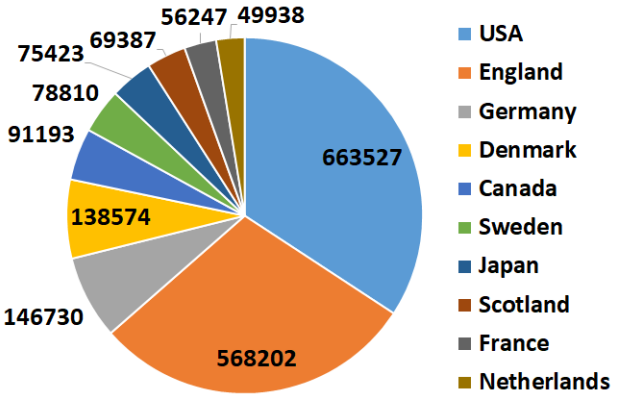

Fig. 3: Country-wise distribution (for the top 10 countries) of spike sequences.

C. Data Visualization

To see if there is any (hidden) clustering in the data, we mapped the data to 2D real vectors using the t-distributed

[3]https://github.com/sarwanpasha/Spike2Vec
[4]https://drive.google.com/drive/folders/1-YmIM8ipFpj-glr9hSF3t6VuofrpgWUa?usp=sharing
[5]https://www.gisaid.org/

Pango Lineage	Region	Labels	Num. Mutations S-gene/Genome	Num. of sequences
B.1.1.7	UK [7]	Alpha	8/17	976077
B.1.351	South Africa [7]	Beta	9/21	20829
B.1.617.2	India [13]	Delta	8/17	242820
P.1	Brazil [12]	Gamma	10/21	56948
B.1.427	California [38]	Epsilon	3/5	17799
AY.4	India [39]	Delta	-	156038
B.1.2	-	-	-	96253
B.1				78741
B.1.177	-	-	-	72298
B.1.1	-		-	44851
B.1.429	-	-	-	38117
AY.12	India [39]	Delta	-	28845
B.1.160	-	-	-	25579
B.1.526	New York [40]	Iota	6/16	25142
B.1.1.519	-	-	-	22509
B.1.1.214	-	-	-	17880
B.1.221	-	-	-	13121
B.1.258	-	-	-	13027
B.1.177.21	-	-	-	13019
D.2	-	-	-	12758
B.1.243	-	-	-	12510
R.1	-	-	-	10034

TABLE I: The SARS-CoV-2 variants which were represented in more than 10,000 sequences (of the $\approx$ 2.5 million sequences). The S/Gen. column represents number of mutations on the Spike (S) gene / entire genome. Total number of amino acid sequences in our dataset is 2,519,386. The variants discussed in this Table comprise 1,995,195 sequences.

stochastic neighbor embedding (t-SNE) approach [41]. Since it was not possible to run the t-SNE algorithm on all $\approx$ 2.5 million spike sequences, we obtained a representative subset of sequences containing 7000 randomly selected sequences such that the proportion of each variant in this subset is equal to its proportion in the original data. The t-SNE plot for Delta, Beta, Iota, Epsilon, and Gamma variants is shown in Figure 4.

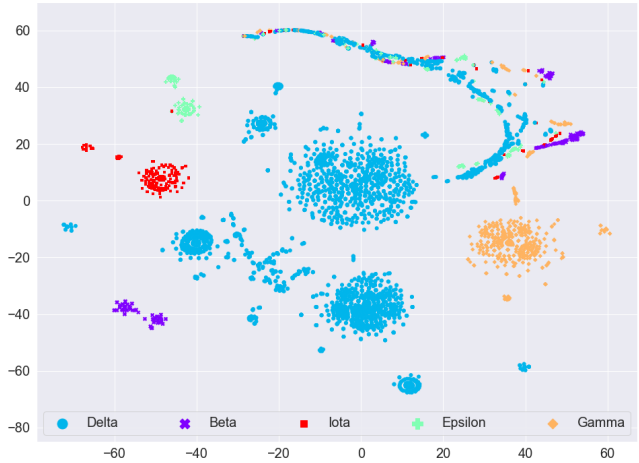

Fig. 4: t-SNE embeddings of spike sequences

1) USA Case Study: Figure 5 shows the COVID-19 spread pattern for three variants in the USA from March 2020 to July 2021. We can see in Figure 5 that after the coronavirus spread hit the peak in April 2021, the number of cases of the coronavirus started decreasing. That was the point where a significant proportion of the population of the USA was vaccinated (hence peak spread reduced).

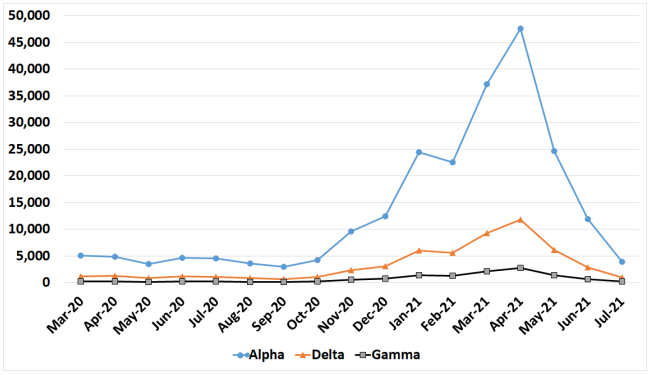

Fig. 5: Spread pattern of Alpha (blue line), Delta (orange line), and Gamma (black line) variants in USA country from March 2020 to July 2021. Y-axis shows the total number of COVID-19 infected patients.

D. Performance Evaluation

To evaluate Spike2vec, we perform classification and clustering on the low dimensional feature vectors that it produces. For the classification task, we use Naive Bayes, Logistic Regression, and Ridge Classifier. For the clustering analysis, we use the K-means algorithm. To evaluate the clustering algorithms, we report accuracy, precision, recall, weighted F1, macro F1, ROC-AUC. We also show the runtime of different classification algorithms. To evaluate the clustering method, we use weighted F1 score.

E. Baseline Algorithm

For the baseline approach, we use the one-hot encoding based approach proposed in [27] (for reference, we call this method as One Hot Embedding "OHE" in rest of the paper). Authors in [27] use a typical one-hot encoding approach to convert the spike sequences into numerical representations. In spike sequences, we have 21 unique amino acids (unique alphabets forming Σ) namely "*ACDEFGHIKLMNPQRSTVWXY*". Also, the length of each spike sequence is 1273 plus an ending character $*$ at the 1274^{th} location. After getting the one-hot encoding based numerical representation for each spike sequence, we will get a feature vector of length $26,733$ corresponding to each spike sequence ($21 \times 1273 = 26,733$). After getting this numerical representation, authors in [27] use the typical principal component analysis (PCA) approach to reduce the dimensionality of the data. In our case, since the size of our data is too huge, simply using PCA would take a lot of computational time. Therefore, in OHE, we use RFF on the one-hot embeddings to get the low dimensional feature vector representations for our baseline model.

Remark 3: If we directly apply the classification algorithms on the one-hot embedding, the underlying classification algorithms simply do not run and exceed the amount of available

Approach	ML Algo.	Acc.	Prec.	Recall	F1 (weighted)	F1 (Macro)	ROC-AUC	Training runtime (sec.)
OHE	NB	0.306370	0.583309	0.306370	0.383127	0.179708	0.594794	566.099
	LR	0.568250	0.498771	0.568250	0.495177	0.196033	0.576567	1309.060
	RC	0.563617	0.479080	0.563617	0.485127	0.174295	0.566908	110.766
Spike2Vec	NB	0.420112	**0.797562**	0.420112	0.521133	0.391974	0.685318	457.5417
	LR	**0.688674**	0.689724	**0.688674**	**0.649236**	**0.490654**	**0.694196**	830.6327
	RC	0.675313	0.680797	0.675313	0.629078	0.447041	0.674559	**95.7315**

TABLE II: Variants Classification Results (1% training set and 99% testing set) for top 22 variants (1995195 spike sequences) discussed in Table I. Best values are shown in bold.

memory (≈3TB). This shows that scalability is the major problem with these typical ML algorithms.

F. Machine Learning Models

For the classification task, we use Naive Bayes (NB), Logistic Regression (LR), and Ridge Classifier (RC) [42] with default parameters. For clustering, we use simple k-means algorithm with default parameters.

Remark 4: For k-means, we use 22 as the number of clusters. We selected the number of clusters using the Elbow method [26]. To do this, we perform clustering with different number of clusters ranging from 2 to 100 and then see the trade-off between the *sum of squared error* (distortion score) and the runtime. After analyzing the trade-off, we use "knee point detection algorithm (KPDA)" [43] to find the optimal value of k for the k-means algorithm.

V. Results and Discussion

In this section, we first present the classification results for Spike2Vec and show that it significantly outperforms the baseline in terms of prediction accuracy. We then show the results for the clustering using weighted F1. In the end, we report the importance of each amino acid by computing information gain.

A. Classification Results

Results for different classification algorithms are shown in Table II. We can observe that overall Logistic regression is almost a clear winner in case of Spike2vec. All the classifiers in case of Spike2Vec clearly outperform the corresponding classifiers with OHE. This performance for different evaluation metrics shows the effectiveness of using k-mers instead of one hot encoding for feature vector representation of the spike sequences. Also, we can observe that although the performance of RC in case of Spike2Vec is not better than LR, it is significantly better than LR and NB however in terms of training runtime. Therefore, we can conclude that overall LR is better in terms of prediction performance with Spike2Vec while RC is better in terms of runtime along with comparable performance to LR.

B. Clustering Results

We also test the performance of Spike2Vec using the k-means clustering method. Results for the clustering methods are shown in Table III. We can observe that Spike2Vec clearly outperforms the baseline models in case of all but 1 variant. The reason for the bad performance in case of Beta and Epsilon variants is due to the fact that they are in comparatively less proportion in the dataset (see Table I). Because of the less information, Spike2Vec is not able to design rich feature vector representation for these variants.

	F1 Score (Weighted) for Different Variants				
Methods	Alpha	Beta	Delta	Gamma	Epsilon
OHE	0.0410	**0.0479**	0.5942	0.6432	0.0571
Spike2Vec	**0.9997**	0.0300	**0.8531**	**0.9680**	**0.2246**

TABLE III: F1 score by applying the k-means clustering algorithm on all 1327 variants (2519386 spike sequences) in the GISAID dataset. Best values are shown in bold.

The contingency table for some of the more well represented variants computed using Spike2Vec is given in Table IV. Since we have a total of 1327 variants, it is not possible to show the contingency table for all of the variants. Therefore, we only show relationship between variants and clusters for some of the popular variants in Table IV.

C. Importance of Each Amino Acid

Authors in [25] compute Information Gain (IG) between the variants and each amino acid (attribute) separately. The main goal for computing IG was to see the importance of each amino acid. Since they are using very small dataset (7000 spike sequences), it is not clear if the information gain values they computed will be the same for our ≈ 2.5 million sequences. Therefore, a "proof of concept" is needed to verify their results (on a larger dataset). The IG is defined as follows:

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

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

where H is the entropy, and p_i is the probability of the class i. We extracted 10 sample datasets of 20000 spike sequences each and computed IG for each of the dataset separately. For each of the dataset, we got same amino acids with the maximum IG values. The values for different amino acids for one of the 10 datasets is given in Figure 6. The Centers for Disease Control and Prevention (CDC) located in USA pointed out mutations at a few positions that take place in different variants [39]. We compare the mutation information from the CDC with the (high) IG values that we got for different amino

	K-means (Cluster IDs)																					
Variant	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
Epsilon	109	0	3186	432	113	219	67	0	160	9	134	0	18	14	0	78	9	3792	113	0	48	41
Alpha	6061	1	175923	23353	5846	11754	3376	0	9466	1041	6889	0	734	1281	0	4160	329	205848	5730	0	3193	2136
Gamma	344	0	10403	1312	327	686	205	0	534	63	390	0	52	77	0	254	14	11977	324	0	182	137
Beta	144	0	3853	436	115	237	64	0	191	8	148	0	19	25	0	81	7	4435	119	0	71	44
Delta	1432	1	43691	5732	1391	2832	831	0	2342	241	1777	0	172	315	0	1016	77	51596	1400	0	836	541

TABLE IV: Contingency tables of variants vs clusters.

acids. According to the CDC, R452L is present in Epsilon and Delta lineages and sub-lineages while K417N, E484K, and N501Y substitutions are present in the Beta variant.

Remark 5: Note that R452L means Amino acid at position 452 was 'R' before and after mutation, it changed to 'L'. Similarly, K417T, E484K, and N501Y substitutions are present in the Gamma variant [39]. We can see in Figure 6 that we obtained the maximum IG values for the same amino acids positions mentioned by CDC. We also made the IG values for all amino acids available online[6].

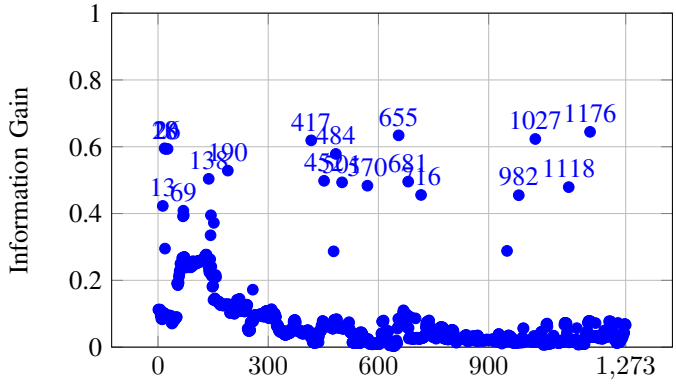

Fig. 6: Information gain for each amino acid position with respect to variants. The x-axis corresponds to amino acid positions in the spike sequences.

VI. Conclusion

We propose an efficient and scalable embedding approach in this paper that can be used to perform different machine learning tasks on the SARS-CoV-2 spike sequences. We show that our model can scale to several million sequences, and it also outperforms the baseline models significantly. Since the COVID-19 disease is relatively new, we do not have enough information available for different coronavirus variants so far. We will explore the new (and existing) variants in more detail in the future. We will also use deep learning models to enhance the prediction performance of Spike2Vec. Using adversarial examples to test the robustness of the ML models in case of spike sequences is another potential future extension.

VII. Acknowledgments

This research was supported by a Georgia State University startup grant.

[6]https://github.com/sarwanpasha/Spike2Vec/blob/main/correlation_data.csv

References

[1] R. Sharma, A. Mateush, and J. Übi, "Tale of three states: analysis of large person-to-person online financial transactions in three baltic countries," in *IEEE BigData*, 2019, pp. 1497–1505.

[2] Y. Dong, D. Yan, A. I. Almudaifer, S. Yan, Z. Jiang, and Y. Zhou, "BELT: A pipeline for stock price prediction using news," in *IEEE BigData*, 2020, pp. 379–410.

[3] J. Kremer, K. Stensbo-Smidt, F. Gieseke, K. S. Pedersen, and C. Igel, "Big universe, big data: Machine learning and image analysis for astronomy," *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 16–22, 2017.

[4] M. Mace, "Comet NEOWISE sizzles as it slides by the sun, providing a treat for observers," *Infrared Processing and Analysis Center*, 2020.

[5] C. K. Leung, Y. Chen, C. S. H. Hoi, S. Shang, and A. Cuzzocrea, "Machine learning and olap on big covid-19 data," in *IEEE BigData*, 2020.

[6] C. K. Leung, Y. Chen, S. Shang, and D. Deng, "Big data science on covid-19 data," in *IEEE BigData*, 2020.

[7] S. Galloway *et al.*, "Emergence of sars-cov-2 b. 1.1. 7 lineage," *Morbidity and Mortality Weekly Report*, vol. 70, no. 3, p. 95, 2021.

[8] S. ZD, L. SY, F. F, C. RH, Z. C, E. MJ *et al.*, "Big data: Astronomical or genomical?" *PLoS Biology*, vol. 13, no. 7, p. e1002195, 2015.

[9] Genome Web, https://www.genomeweb.com/infectious-disease/cdc-commits-90m-create-public-health-pathogen-genomics-research-centers, [Online; accessed 5-September-2021].

[10] Reuters, https://www.reuters.com/article/us-china-genomics-state/chinese-state-fund-invests-in-gene-firm-bgi-idUSKBN2AM0AT, [Online; accessed 5-September-2021].

[11] A. Rambaut, E. C. Holmes, A. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and O. G. Pybus, "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," *Nature Microbiology*, vol. 5, no. 11, pp. 1403–1407, 2020.

[12] F. Naveca *et al.*, "Phylogenetic relationship of sars-cov-2 sequences from amazonas with emerging brazilian variants harboring mutations e484k and n501y in the spike protein," *Virological. org*, vol. 1, 2021.

[13] P. Yadav *et al.*, "Neutralization potential of covishield vaccinated individuals sera against b. 1.617. 1," *bioRxiv*, vol. 1, 2021.

[14] E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg, W. R. Hinsley, D. J. Laydon, G. Dabrera, Á. O'Toole *et al.*, "Assessing transmissibility of sars-cov-2 lineage b. 1.1. 7 in england," *Nature*, vol. 593, pp. 266–269, 2021.

[15] M. McCallum, J. Bassi, A. Marco, A. Chen, A. Walls, J. Iulio, M. Tortorici, M. Navarro, C. Silacci-Fregni, C. Saliba, M. Agostini, D. Pinto, K. Culap, S. Bianchi, S. Jaconi, E. Cameroni, J. Bowen, S. Tilles, M. Pizzuto, S. Guastalla, G. Bona, A. Pellanda, C. Garzoni, W. Van Voorhis, L. Rosen, G. Snell, A. Telenti, H. Virgin, L. Piccoli, D. Corti, and D. Veesler, "Sars-cov-2 immune evasion by variant b.1.427/b.1.429," vol. 1, 2021, doi:10.1101/2021.03.31.437925.

[16] Y. Furuse, "Genomic sequencing effort for sars-cov-2 by country during the pandemic," *International Journal of Infectious Diseases*, vol. 103, pp. 305–307, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1201971220325571

[17] A. Maxmen, "Why US coronavirus tracking can't keep up with variants," *Nature*, vol. 592, 2021.

[18] GISAID Website, https://www.gisaid.org/, [Online; accessed 5-September-2021].

[19] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Z.-G. Song, Y. Hu, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei *et al.*, "A new coronavirus associated with human respiratory disease in china," *Nature*, vol. 579, pp. 265–269, 2020.

[20] F. Wu, A. Xiao, J. Zhang, K. Moniz, N. Endo, F. Armas, M. Bushman, P. R. Chai, C. Duvallet, T. B. Erickson, K. Foppe, N. Ghaeli, X. Gu, W. P. Hanage, K. H. Huang, W. L. Lee, K. A. McElroy, S. F. Rhode,

M. Matus, S. Wuertz, J. Thompson, and E. J. Alm, "Wastewater surveillance of sars-cov-2 across 40 u.s. states from february to june 2020," *Water Research*, vol. 202, p. 117400, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0043135421005984

[21] J. Gardy and N. Loman, "Towards a genomics-informed, real-time, global pathogen surveillance system," *Nature Reviews Genetics*, vol. 19, pp. 9–20, 2018.

[22] F. Li, "Structure, function, and evolution of coronavirus spike proteins," *Annual Review Virology*, vol. 3, no. 1, pp. 237–261, 2016.

[23] A. Walls, Y. Park, and M. Tortorici, "Structure, function and antigenicity of the sars-cov-2 spike glycoprotein," *Cell*, vol. 181, no. 2, pp. 281–292, 2020.

[24] K. S. Corbett, D. K. Edwards, S. R. Leist *et al.*, *Nature*, vol. 586, pp. 567–571, 2020.

[25] S. Ali, B. Sahoo, N. Ullah, A. Zelikovskiy, M. Patterson, and I. Khan, "A k-mer based approach for sars-cov-2 variant identification," *To Appear at: International Symposium on Bioinformatics Research and Applications (ISBRA)*, 2021.

[26] S. Ali, T. e Ali, M. A. Khan, I. Khan, and M. Patterson, "Effective and scalable clustering of sars-cov-2 sequences," *To appear at: International Conference on Big Data Research (ICBDR)*, 2021.

[27] K. Kuzmin *et al.*, "Machine learning methods accurately predict host specificity of coronaviruses based on spike sequences alone," *Biochemical and Biophysical Research Communications*, vol. 533, pp. 553–558, 2020.

[28] S. Ali, S. Ciccolella, L. Lucarella, G. Della Vedova, and M. D. Patterson, "Simpler and faster development of tumor phylogeny pipelines," *To appear at: Journal of Computational Biology (JCB)*, 2021.

[29] S. Ali, H. Mansoor, N. Arshad, and I. Khan, "Short term load forecasting using smart meter data," in *International Conference on Future Energy Systems (e-Energy)*, 2019, pp. 419–421.

[30] A. Melnyk, F. Mohebbi, S. Knyazev, B. Sahoo, R. Hosseini, P. Skums, A. Zelikovskiy, and M. D. Patterson, "From alpha to zeta: Identifying variants and subtypes of sars-cov-2 via clustering," *To appear at: Journal of Computational Biology (JCB)*, 2021.

[31] S. Knyazev, V. Tsyvina, A. Shankar, A. Melnyk, A. Artyomenko, T. Malygina, Y. B. Porozov, E. M. Campbell, S. Mangul, W. M. Switzer *et al.*, "Accurate assembly of minority viral haplotypes from next-generation sequencing through efficient noise reduction," *Nucleic Acids Research*, p. 264242, 2020.

[32] M. Farhan, J. Tariq, A. Zaman, M. Shabbir, and I. Khan, "Efficient approximation algorithms for strings kernel based sequence classification," in *Advances in neural information processing systems (NeurIPS)*. ., 2017, pp. 6935–6945.

[33] A. Rahimi, B. Recht *et al.*, "Random features for large-scale kernel machines." in *NIPS*, vol. 3, no. 4, 2007, p. 5.

[34] H. Peng, N. Pappas, D. Yogatama, R. Schwartz, N. A. Smith, and L. Kong, "Random feature attention," in *International Conference on Learning Representations (ICLR)*, 2021.

[35] S. Ali, "Cache replacement algorithm," *arXiv preprint arXiv:2107.14646*, 2021.

[36] R. Rizzi, S. Beretta, M. Patterson, Y. Pirola, M. Previtali, G. D. Vedova, and P. Bonizzoni, "Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era," *Quantitative Biology*, vol. 7, no. 4, pp. 278–292, 2019.

[37] P. Devijver and J. Kittler, "Pattern recognition: A statistical approach," in *London, GB: Prentice-Hall*, 1982, pp. 1–448.

[38] W. Zhang *et al.*, "Emergence of a novel sars-cov-2 variant in southern california," *Jama*, vol. 325, no. 13, pp. 1324–1326, 2021.

[39] SARS-CoV-2 Variant Classifications and Definitions, https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html, [Online; accessed 5-September-2021].

[40] A. West Jr *et al.*, "Detection and characterization of the sars-cov-2 lineage b. 1.526 in new york," *bioRxiv*, 2021.

[41] L. Van der M. and G. Hinton, "Visualizing data using t-sne." *Journal of Machine Learning Research (JMLR)*, vol. 9, no. 11, 2008.

[42] A. Singh, B. S. Prakash, and K. Chandrasekaran, "A comparison of linear discriminant analysis and ridge classifier on twitter data," in *2016 International Conference on Computing, Communication and Automation (ICCCA)*, 2016, pp. 133–138.

[43] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a" kneedle" in a haystack: Detecting knee points in system behavior," in *International conference on distributed computing systems workshops*. IEEE, 2011, pp. 166–171.