Accurate Identification of SARS-CoV-2 from Viral Genome Sequences using Deep Learning

Alejandro Lopez-Rincon*[1], Alberto Tonda[2], Lucero Mendoza-Maldonado[3], Eric Claassen[5], Johan Garssen[1,4] and Aletta D. Kraneveld[1]

[1]Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, Universiteitsweg 99, 3584 CG Utrecht, the Netherlands; [2]INRAe UMR 518 MIA-Paris, c/o 113 rue Nationale, 75103, Paris, France; [3] Centro Universitario de Ciencias de la Salud, Universidad de Guadalajara, Sierra Mojada No. 950, Col. Independencia C.P. 44348 Guadalajara, Jalisco, Mexico; [4]Department Immunology, Danone Nutricia research, Uppsalalaan 12, 3584 CT Utrecht, the Netherlands; [5]Athena Institute, Vrije Universiteit, De Boelelaan 1085, 1081 HV Amsterdam, the Netherlands.

Abstract

One of the reasons for the fast spread of SARS-CoV-2 is the lack of accuracy in detection tools in the clinical field. Molecular techniques, such as quantitative real-time RT-PCR and nucleic acid sequencing methods, are widely used to identify pathogens. For this particular virus, however, they have an overall unsatisfying detection rate, due to its relatively recent emergence and still not completely understood features. In addition, SARS-CoV-2 is remarkably similar to other Coronaviruses, and it can present with other respiratory infections, making identification even harder. To tackle this issue, we propose an assisted detection test, combining molecular testing with deep learning. The proposed approach employs a state-of-the-art deep convolutional neural network, able to automatically create features starting from the genome sequence of the virus. Experiments on data from the Novel Coronavirus Resource (2019nCoVR) show that the proposed approach is able to correctly classify SARS-CoV-2, distinguishing it from other coronavirus strains, such as MERS-CoV, HCoV-NL63, HCoV-OC43, HCoV-229E, HCoV-HKU1, and SARS-CoV regardless of missing information and errors in sequencing (noise). From a dataset of 553 complete genome non-repeated sequences that vary from 1,260 to 31,029 bps in length, the proposed approach classifies the different coronaviruses with an average ac-

curacy of 98.75% in a 10-fold cross-validation, identifying SARS-CoV-2 with an AUC of 98%, specificity of 0.9939 and sensitivity of 1.00 in a binary classification. Then, using the same basis, we classify SARS-CoV-2 from 384 complete viral genome sequences with human host, that contain the gene *ORF1ab* from the NCBI with a 10-fold accuracy of 98.17% , a specificity of 0.9797 and sensitivity of 1.00. Furthermore, an in-depth analysis of the results allow us to identify base pairs sequences that are unique to SARS-CoV-2 and do not appear in other virus strains, that could then be used as a base for designing new primers and examined by experts to extract further insights. These preliminary results seem encouraging enough to identify deep learning as a promising research venue to develop assisted detection tests for SARS-CoV-2. At this end the interaction between viromics and *deep learning*, will hopefully help to solve global infection problems. In addition, we offer our code and processed data to be used for diagnostic purposes by medical doctors, virologists and scientists involved in solving the SARS-CoV-2 pandemic. As more data become available we will update our system.

Keywords: convolutional neural networks, coronavirus, deep learning, SARS-CoV-2

1. Introduction

The Coronaviridae family presents a positive sense, single-strand RNA genome. This viruses have been identified in avian and mammal hosts, including humans. Coronaviruses have genomes from 26.4 kilo base-pairs (kbps) to 31.7 kbps, with G + C contents varying from 32% to 43%, and human-infecting coronaviruses include SARS-CoV, MERS-CoV, HCoV-OC43, HCoV-229E, HCoV-NL63 and HCoV-HKU1 [1]. In December 2019, SARS-CoV-2, a novel, human-infecting Coronavirus was identified in Wuhan, China, using Next Generation Sequencing [2].

As a typical RNA virus, new mutations appears every replication cycle of Coronavirus, and its average evolutionary rate is roughly 10-4 nucleotide sub-

stitutions per site each year [2]. In the specific case of SARS-CoV-2, RT-qPCR testing using primers in ORF1ab and N genes have been used to identified the infection in humans. However, this method presents a high false negative rate (FNR), with a detection rate of 30-50% [3, 4]. This low detection rate can be explained by the variation of viral RNA sequences within virus species, and the viral load in different anatomic sites [5]. Population mutation frequency of site 8,872 located in ORF1ab gene and site 28,144 located in ORF8 gene gradually increased from 0 to 29% as the epidemic progressed [6].

As of March 6^{th} of 2020, the new SARS-CoV-2 has 98,192 confirmed cases across 88 countries, with 17,481 cases outside of China [7]. In addition, SARS-CoV-2 has an estimated mortality rate of 3-4%, and it is spreading faster than SARS-CoV and MERS-CoV [8]. SARS-CoV-2 assays can yield false positives if they are not targeted specifically to SARS-CoV-2, as the virus is closely related to other Coronavirus organisms. In addition, SARS-CoV-2 may present with other respiratory infections, which make it even more difficult to identify [9, 10]. Thus, it is fundamental to improve existing diagnostic tools to contain the spread. For example, diagnostic tools combining computed tomography (CT) scans with deep learning have been proposed, achieving an improved detection accuracy of 82.9% [11]. Another solution for identifying SARS-CoV-2 is additional sequencing of the viral complementary DNA (cDNA). We can use sequencing data with cDNA, resulting from the PCR of the original viral RNA; e,g, Real-Time PCR amplicons (Fig. 1) to identify the SARS-CoV-2 [12].

Classification using viral sequencing techniques is mainly based on alignment methods such as FASTA [13] and BLAST [14]. These methods rely on the assumption that DNA sequences share common features, and their order prevails among different sequences [15, 16]. However, these methods suffer from the necessity of needing base sequences for the detection [17]. Nevertheless, it is necessary to develop innovative improved diagnostic tools that target the genome to improve the identification of pathogenic variants, as sometimes several tests, are needed to have an accurate diagnosis. As an alternative deep learning methods have been suggested for classification of DNA sequences, as these methods

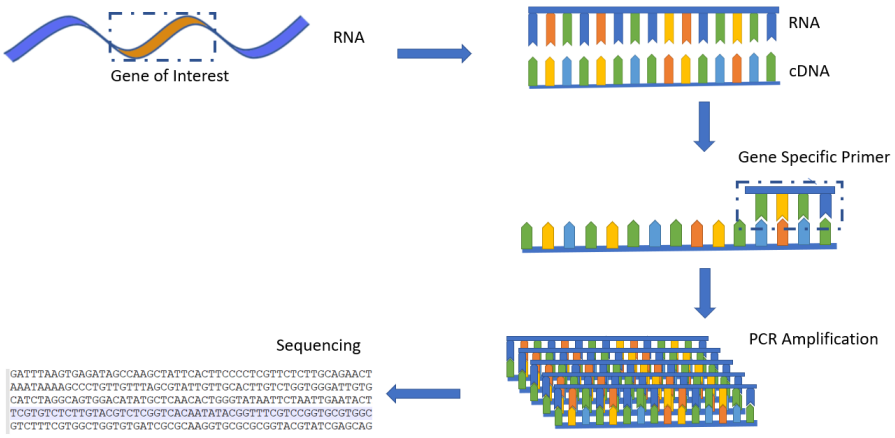


Figure 1: PCR Amplicons sequencing procedure.

do not need pre-selected features to identify or classify DNA sequences. Deep Learning has been efficiently used for classification of DNA sequences, using one-hot label encoding and Convolution Neural Networks (CNN) [18, 19], albeit the examples in literature are featuring DNA sequences of length up to 500 bps, only.

In particular, for the case of viruses, Next Generation Sequencing (NGS) genomic samples might not be identified by BLAST, as there are no reference sequences valid for all genomes, as viruses have high mutation frequency [20]. Alternative solutions based on deep learning have been proposed to classify viruses, by dividing sequences into pieces of fixed lengths, from 300 bps [20] to 3,000 bps [21]. However, this approach has the negative effect of potentially ignoring part of the information contained in the input sequence, that is disregarded if it cannot completely fill a piece of fixed size.

Given the impact of the world-wide outbreak, international efforts have been made to simplify the access to viral genomic data and metadata through international repositories, such as; the 2019 Novel Coronavirus Resource (2019nCoVR) repository [6] and the National Center for Biotechnology Information (NCBI) [22], expecting that the easiness to acquire information would make it possible to de-

velop medical countermeasures to control the disease worldwide, as it happened in similar cases earlier [23, 24, 25]. Thus, taking advantage of the available information of international resources without any political and/or economic borders, we propose an innovative system based on viral gene sequencing.

Differently from previous works in literature, that use of deep learning with fixed length features and one-hot label encoding, in this work we propose the use of a different encoding to input the full sequence as a whole. In addition, we use as base input 31,029 as an input vector, which is the maximum length of available DNA sequences for Coronavirus. Finally, we propose a novel architecture for the deep network, inspired by successful applications in cancer detection starting from miRNA [26].

2. Methods

2.1. Data

2.1.1. Classification of Coronaviruses

SARS-CoV-2 identification can give wrong results, as the virus is difficult to distinguish from other Coronaviruses, due to their genetic similarity. In addition, people with SARS-CoV-2 may present other infections besides the virus [9, 10]. Therefore, it is important to be able to properly classify SARS-CoV-2 from other Coronaviruses.

From the repository 2019 Novel Coronavirus Resource (2019nCoVR) [6], we downloaded all the available sequences with the query *Nucleotide Completeness="complete" AND host="homo sapiens"*, for a total of 588 samples. Next, we removed all repeated sequences, resulting in 553 unique sequences of variable length (1,260-31,029 bps). The data was organized and labeled as summarized by Table 1. We grouped HCoV-229E and HCoV-OC43 in the same class, as they are mostly known as Coronaviruses responsible for the common cold [27]; the two available samples of HCoV-4408 were also added to the same class, as it is a Betacoronavirus 1, as HCoV-OC43. In a similar fashion, we grouped HCoV-NL63 and HCoV-HKU1, as they are both associated with acute respiratory

infections (ARI) [28]. Finally, we grouped SARS-CoV/SARS-CoV-P2/SARS-CoV HKU-39849 [29]/SARS-CoV GDH-BJH01 organisms together, as they are all strains of SARS.

Table 1: Organism, assigned label, and number of samples in the unique sequences obtained from the repository [6]. We use the NCBI organism naming convention [30].

Organism	Label	Number of Samples
SARS-CoV-2	0	66
MERS-CoV	1	240
HCoV-OC43	2	140
HCoV-229E	2	22
HCoV-4408	2	2
HCoV-NL63	3	58
HCoV-HKU1	3	17
SARS-CoV	4	7
SARS-CoV P2	4	1
SARS-CoV HKU-39849	4	1
SARS-CoV GDH-BJH01	4	1

To encode the cDNA data into an input tensor for the CNN, we assigned numeric values to the different bases; C=0.25, T=0.50, G=0.75, A=1.0 (see Fig. 2). All missing entries were assigned the value 0.0. This procedure is different from previous methods, that relied upon one-hot encoding [21, 20], and has the advantages of making the input more human-readable and do not multiply the amount of memory required to store the information. We divide the available samples in two parts, 90% for training and validation (80% training, 10% validation), and 10% for testing, in a 10-fold cross-validation scheme. k-fold cross-validation is a procedure by which available data is divided into k parts, called *folds*. At each iteration i, the i-th fold is used as a test set, while all the other folds are used as training. At the end of the k-th iteration, the average performance of the model in test over all folds provides a good estimate of the generality of the results. In this particular case, we use stratified folds, that preserve the same proportion of classes in every fold. The procedure is summarized by Fig. 3.

Figure 2: Coding for the input sequences.

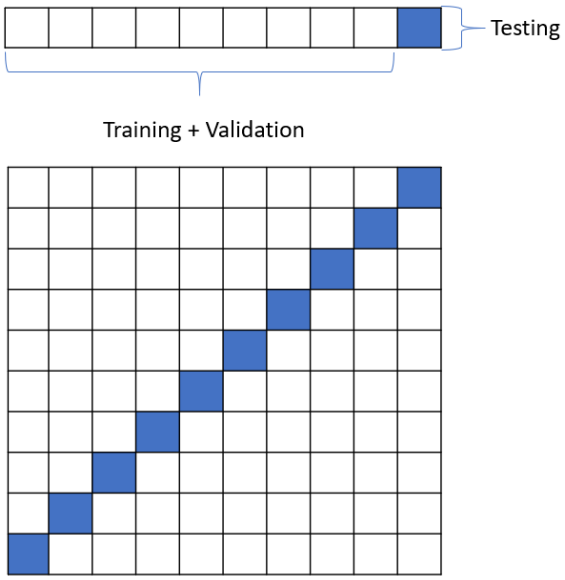


Figure 3: Scheme of a k-fold cross-validation. Available data is divided into k parts. At each iteration i, the i-th fold is used for testing, while all the others are used as a training set.

2.1.2. Separating SARS-CoV-2 from other viruses containing gene ORF1ab

Two thirds of the Coronaviruses' genome contain the ORF1ab gene [1]. Therefore, it is important that we are able to differentiate SARS-CoV-2 from similar viruses, like Astroviruses. From the NCBI repository [30], we downloaded the genome sequences corresponding to the following search: *gene="ORF1ab" AND host="homo sapiens" AND "complete genome"*. This resulted in 402 sequences, distributed as described in Table 2. For this data, we assigned SARS-CoV-2 label 0, and grouped the rest of the organisms together in label 1. Next, we removed all the repeated sequences, obtaining a total of 384 unique sequences, with 45 samples belonging to SARS-CoV-2. The genomic data was translated to digits using the encoding previously described in Subsection 2.1.1.

Table 2: Organism, assigned label, and number of samples in the unique sequences obtained from the repository NCBI [30].

Virus	Label	Number of Samples
SARS-CoV-2	0	45
MERS-CoV	1	180
HCoV-OC43	1	105
HCoV-NL63	1	29
HCoV-HKU1	1	13
HCoV-4408	1	2
HCoV-229E	1	3
HCoV-EMC	1	3
HAstV-VA1	1	1
HAstV-BF34	1	1
HMO-A	1	1
HAstV-SG	1	1

2.2. Convolutional Neural Network

The deep learning model used for the experiments is a CNN with a convolutional layer with max pooling, a fully connected layer, and a final softmax layer, as described in Fig. 4. The input is a vector of 31,029 elements,

which is the maximum size of the genome sequences in the dataset. The convolutional layer is characterized by three hyperparameters, as shown in Fig. 5: $w = 12, wd = 21, h = 148$. The fully connected layer has 196 ReLu units and it is set with a dropout probability of $p_d = 0.5$ during training, to improve generality; moreover, a $l2$ regularization is applied to the cross-categorical entropy loss function, considering all weights in the convolutional layer, with $\beta = 10^{-3}$. The optimizer used for the weights is Adaptive Moment Estimation (Adam) [31], with learning rate $lr = 10^{-5}$, run for 500 epochs. The hyper-parameters used in the experiments were selected after a set of preliminary trials. All the necessary code was developed in Python 3, using the `tensorflow` [32] and `keras` [33] libraries for deep learning, and has been made available on an open GitHub repository[1].

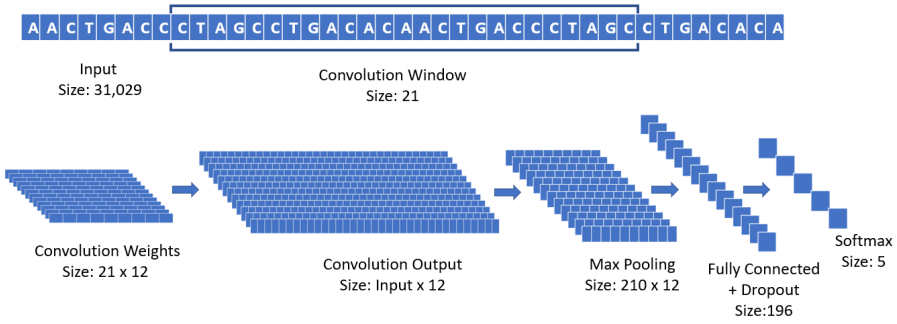


Figure 4: Architecture of the deep convolutional neural network used in the experiments.

3. Results

3.1. Classification of SARS-CoV-2 among Coronaviruses

In the first test, we separated the SARS-CoV-2 from other sequences available at the repository 2019 Novel Coronavirus Resource (2019nCoVR) [6]. We obtained a 10-fold mean test accuracy of $\mu = 0.9875$ with $\sigma = 0.0160$. The

[1] `https://github.com/albertotonda/deep-learning-coronavirus-genome`

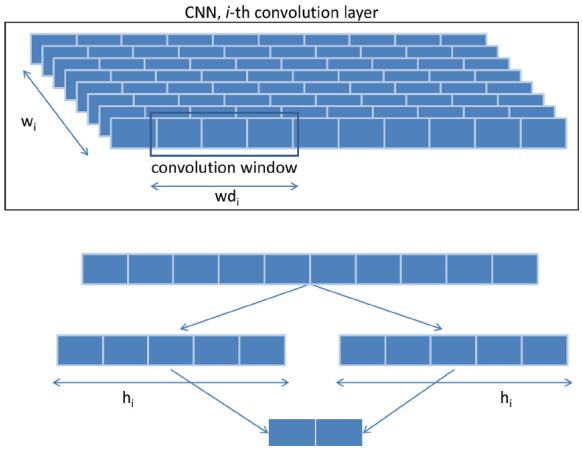


Figure 5: Structure of a convolutional layer in the network. For each of the $w = 12$ filters, the convolution window of size $wd = 21$ is slid over the data, one step at a time. The results of the convolution are then passed through a max pooling layer of size $h = 148$, that helps making the representation approximately invariant to small translations of the input.

resulting confusion matrix (Fig. 6) shows that only 3 out of the 66 SARS-CoV-2 sequences were mistakenly assigned to another class. The binarized curve of the test (Fig. 7) has an area under the curve (AUC) of 0.98, with a specificity of 0.9939 and sensitivity of 1.00. This is considered an outstanding performance, according to the guidelines provided by [34, 35].

As viruses are characterized by high mutation frequencies, to assess the robustness of our approach, we performed further experiments where we added noise to the dataset, simulating possible future mutations. 5% noise was added by randomly selecting 1,551 positions from each sequence, from the 31,029 available, and modifying each selected base to another, or to a missing value, randomly. A new 10-fold cross-validation classification run on the noisy dataset yields an average accuracy $\mu = 0.9674$ with a $\sigma = 0.0158$. Figs. 8 and 9 show the resulting confusion matrix and ROC curve, respectively. This gives a AUC of 0.97, with a specificity of 0.9939 and sensitivity of 0.90.

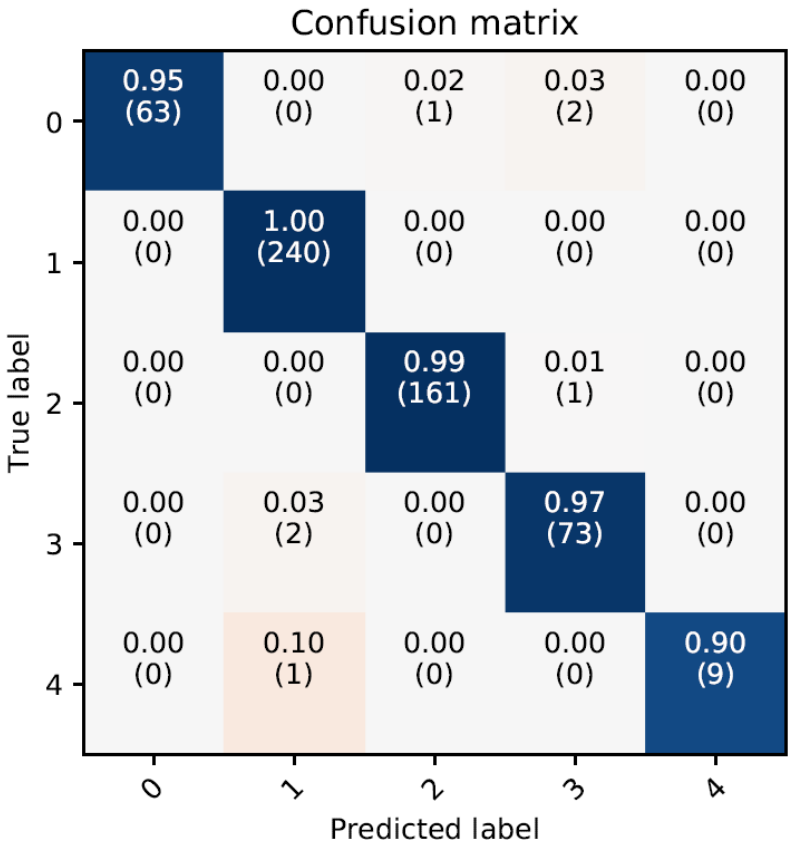


Figure 6: Confusion matrix resulting from the test of a 10-fold cross-validation, comprising 553 samples belonging to 5 different classes.

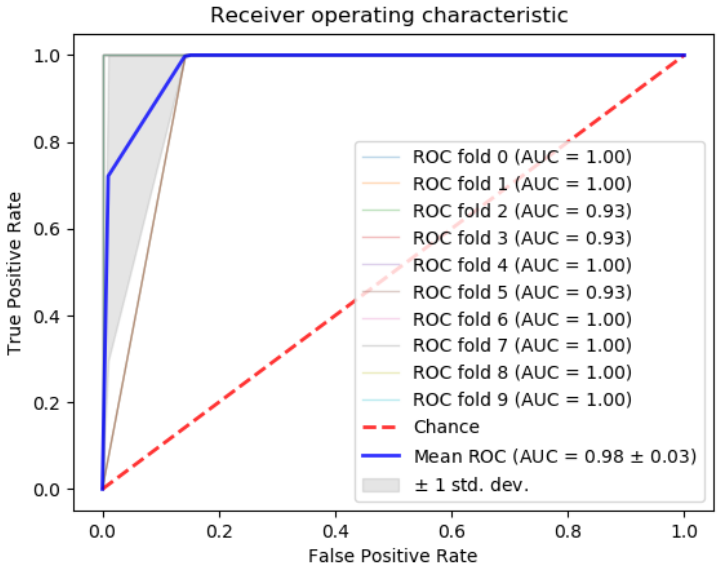


Figure 7: Binarized ROC curve of the 553 sequences, where we consider samples belonging to SARS-CoV-2 as class 0, and all the rest as class 1.

3.2. Separating SARS-CoV-2 from other viruses containing gene ORF1ab

In a next batch of experiments, we aim to distinguish SARS-CoV-2 from other genome sequences from NCBI [30], with the following search parameters: *gene="ORF1ab" AND host="homo sapiens" AND "complete genome"*. We get a 10-fold average accuracy of $\mu = 0.9817$ with a $\sigma = 0.0167$. The resulting confusion matrix (Fig. 6) shows that 7 out of the 45 SARS-CoV-2 sequences, were classified in another class. The ROC curve of the test (Fig. 11) has an area under the curve (AUC) of 0.92 , with a specificity of 0.9797 and sensitivity of 1.00.

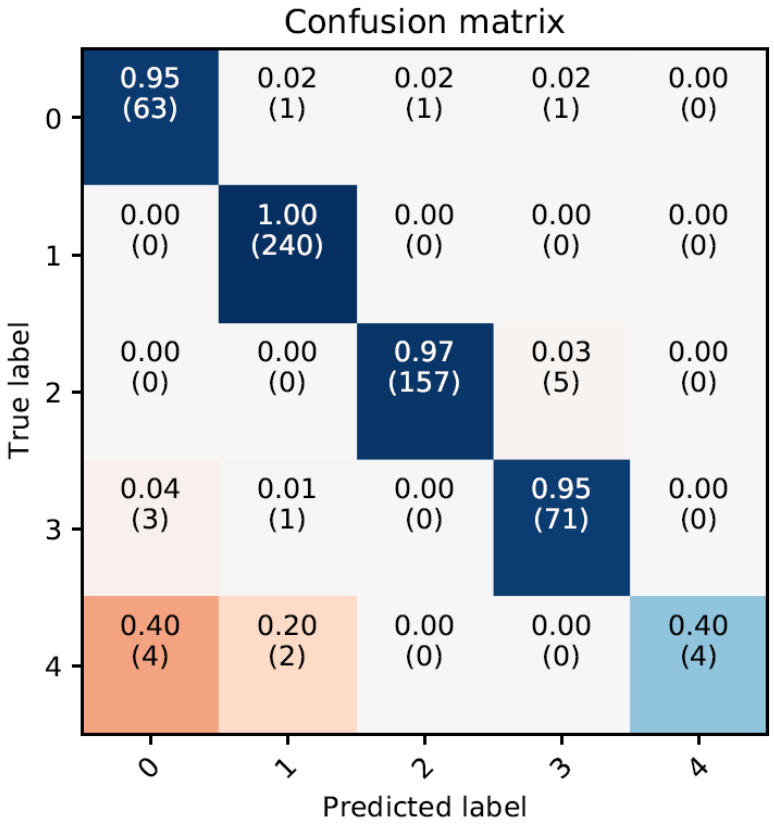


Figure 8: Confusion matrix resulting from the test of a 10-fold cross-validation, comprising 553 samples belonging to 5 different classes, with a 5% noise in the dataset.

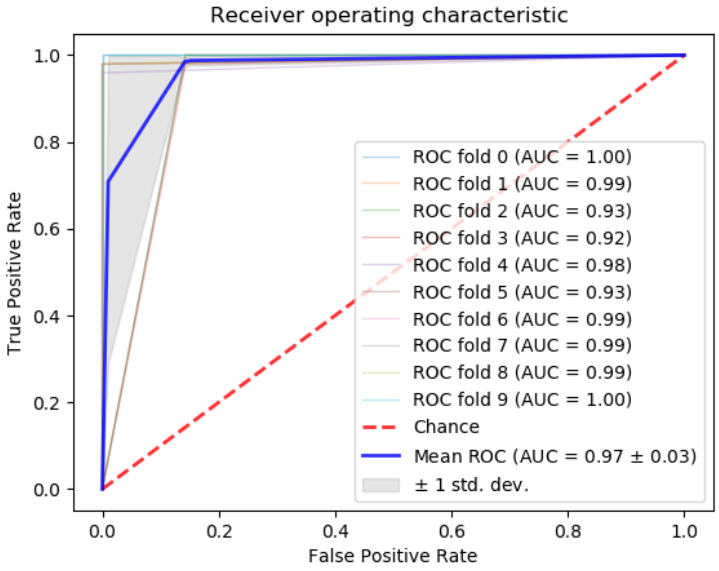


Figure 9: Binarized ROC curve of the 553 sequences, where we consider samples belonging to SARS-CoV-2 as class 0, and all the rest as class 1, with 5% added noise.

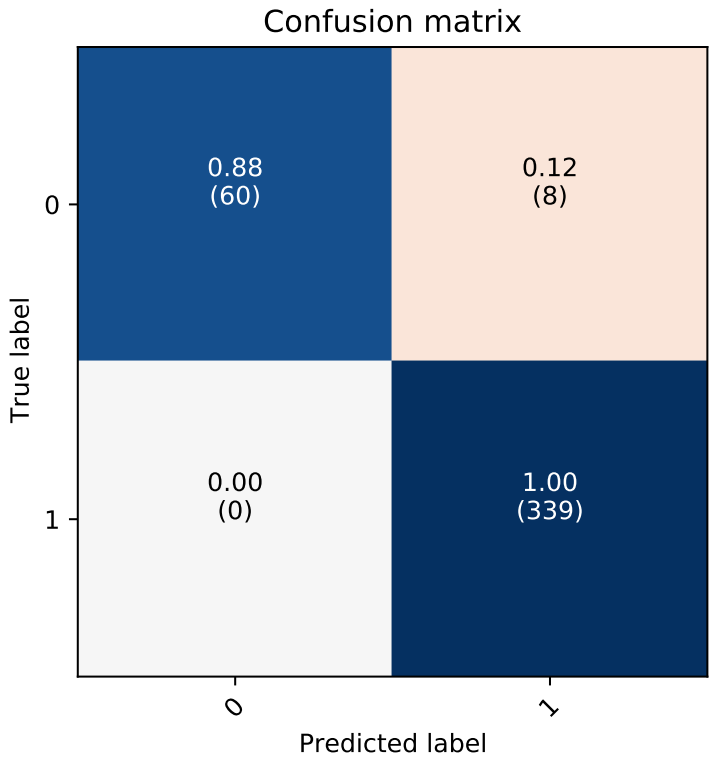


Figure 10: Confusion Matrix of the proposed approach on the 384 NCBI sequences, binarizing the problem with only two classes. Label 0 corresponds to SARS-CoV-2, label 1 to all the other virus strains.

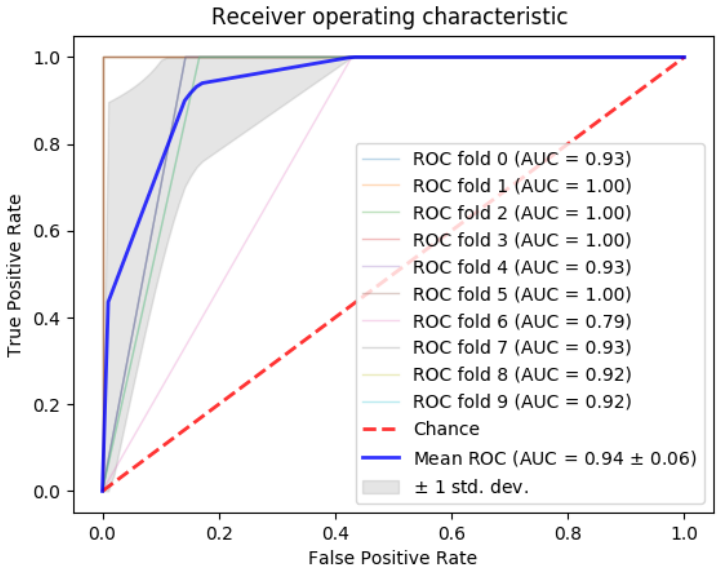


Figure 11: ROC curve of the proposed approach classifying the 384 NCBI sequences, where we consider SARS-CoV-2, as class 0 and the rest as 1.

4. Feature Detection

The convolutional layers of CNNs *de-facto* learn new features to characterize the problem, directly from the data. In this specific case, the new features are specific sequences of base pairs that can more easily separate different virus strains (Fig. 12). By analyzing the result of each filter in a convolutional layer, and how its output interacts with the corresponding max pooling layer, it is possible to detect human-readable sequences of base pairs that might provide domain experts with important information. It is important to notice that these sequences are not bound to specific locations of the genome; thanks to its structure, the CNN is able to detect them and recognize their importance even if their position is displaced in different samples.

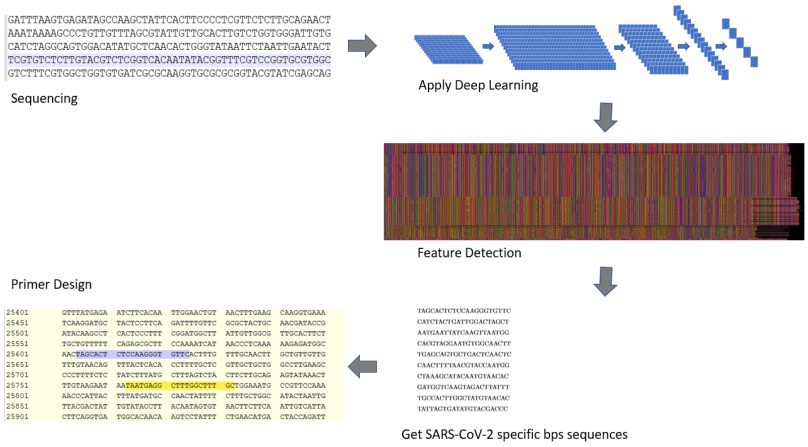
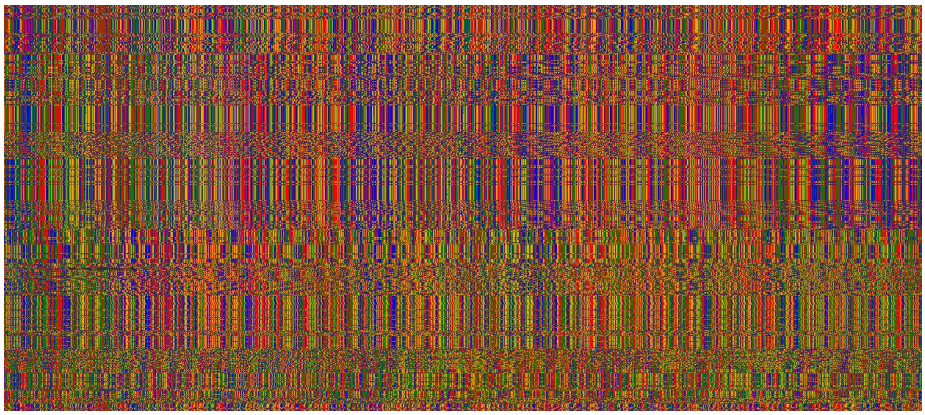


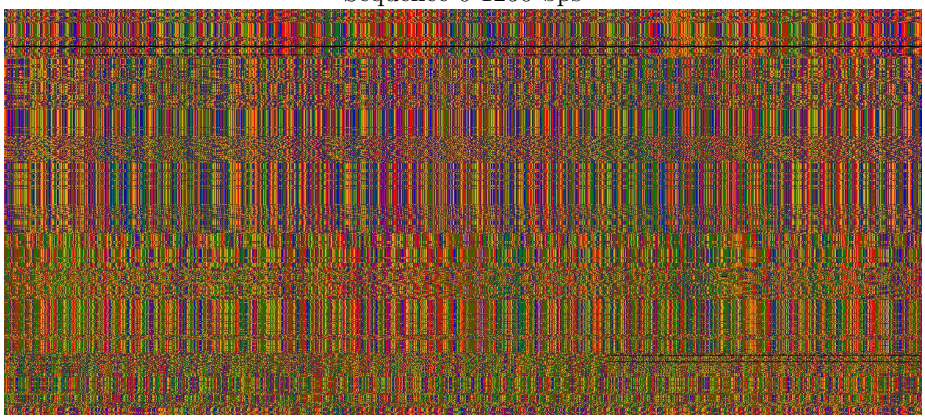
Figure 12: Overall procedure to find the specific SARS-CoV-2 sequences.

For this purpose, we use the trained CNN described in Subsection 2.2, that obtained an accuracy of 98.75% in a 10-fold cross-validation. In a first step, we plot the inputs and outputs of the convolutional layer, to visually inspect for patterns. As an example, in Fig. 13 we report the visualization of the first 2,500 bps of each of the 553 samples considered in the first experiment. Each filter slides a 21-bps window over the input, and for each step produces a single value. The output of a filter is thus a sequence of values in $(0, 1)$: Fig. 14,15

shows the outputs of four out of the twelve filters in the CNN, for the first 2,500 bps of all 553 samples, mapped to a monochromatic image where the closest a value is to 1, the whiter the corresponding pixel is.



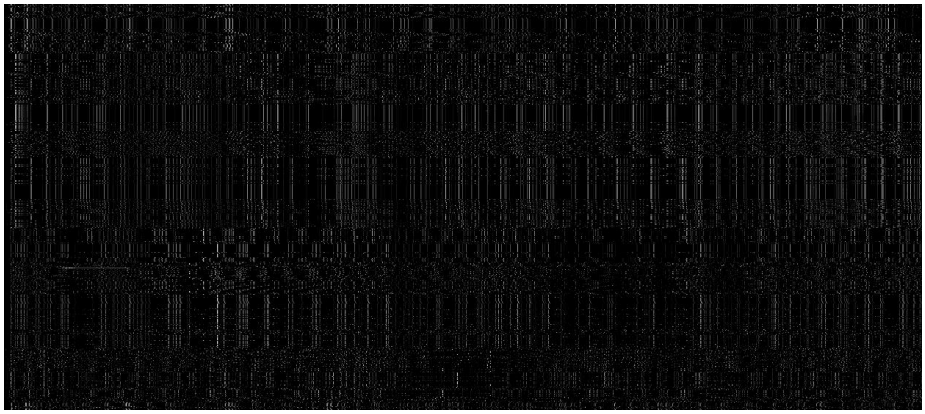
Sequence 0-1250 bps



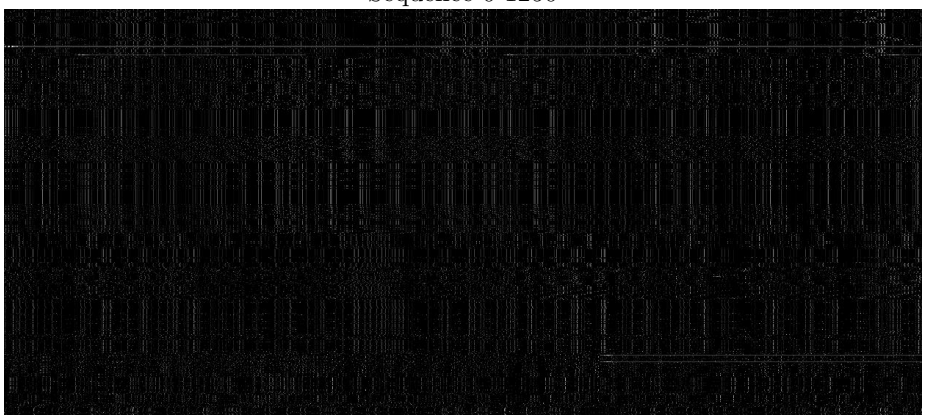
Sequence 1250-2500 bps

Figure 13: cDNA visualization for the first 2,500 bps from the input dataset, for each of the 553 samples. Each sample is represented by a horizontal line of pixels. Colored pixels represent bases: G=green, C=blue, A=red, T=orange, missing=black.

The output of the max pooling layer for each filter is then further inspected for patterns. An example of the output of the max pooling layer for the first two filters is displayed in Fig. 16: it is noticeable how the different classes can be already visually distinguished. At this step, we identify filter 1 as the most

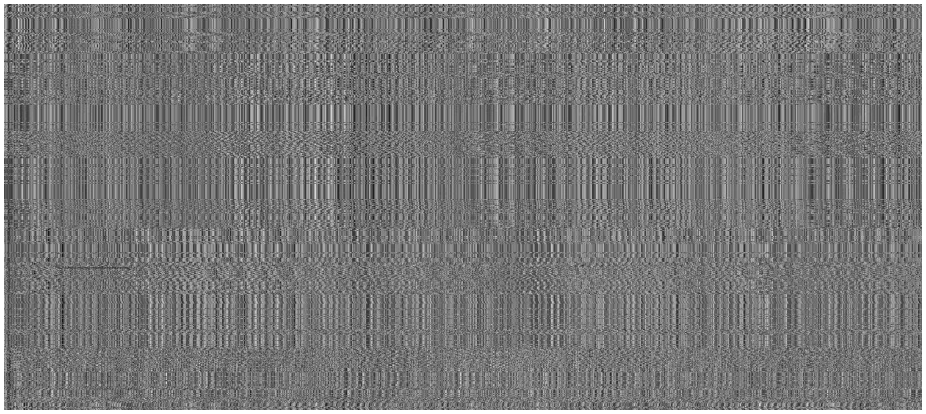


Sequence 0-1250

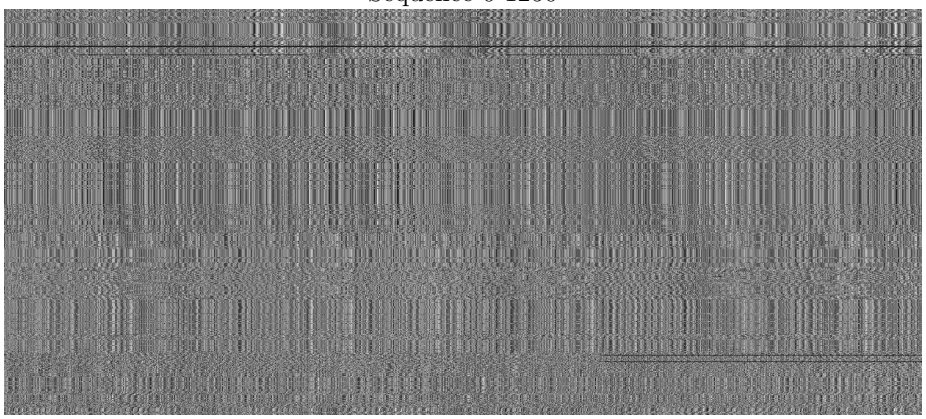


Sequence 1250-2500

Figure 14: The output of convolutional filter 0, for the input given in Fig. 13. The output of the filters is a series of continuous values in $(0, 1)$, here represented in grayscale, with higher values closer to white.



Sequence 0-1250



Sequence 1250-2500

Figure 15: The output of convolutional filter 1, for the input given in Fig. 13. The output of the filters is a series of continuous values in $(0, 1)$, here represented in grayscale, with higher values closer to white

promising, as it seems to focus on the a few relevant points in the genome, and it is thus most likely able to identify meaningful sequences.

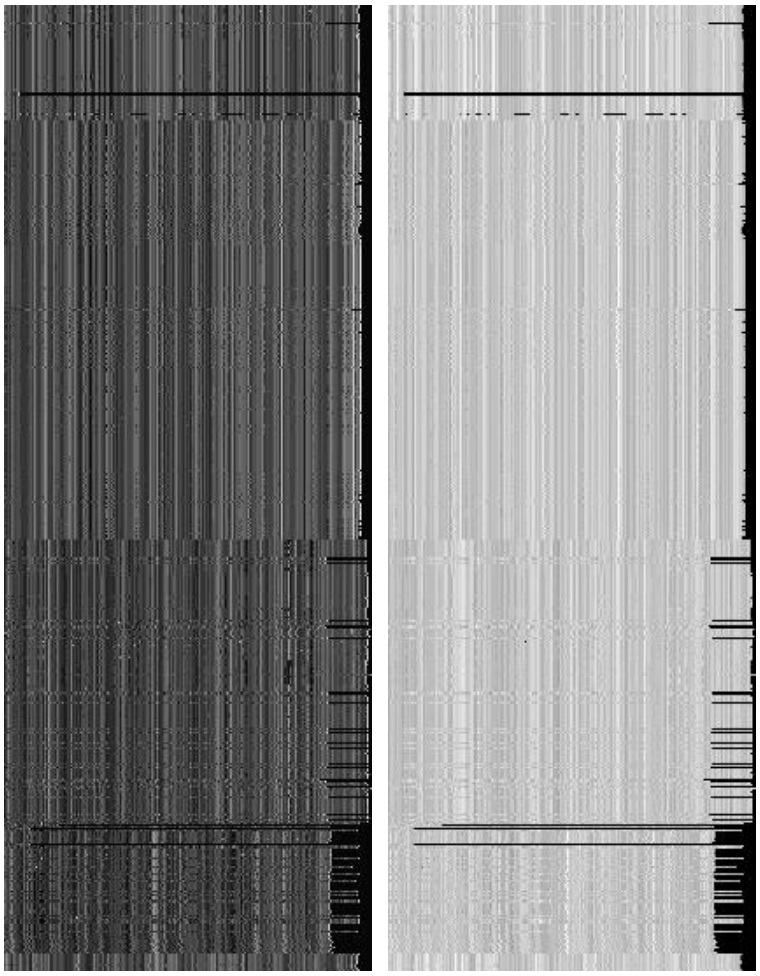


Figure 16: Visualization of the output of the max pooling for the first two filters of the CNN, with the data from the convolutional filters (Fig. 14,15) in input. Different patterns for samples from different classes are recognizable from a simple visual inspection.

190 Given this data, it is now possible to identify the 21-bps sequences (created by the first convolutional filter) that obtained the highest output values in the max pooling layer of filter 1, in a section of 148 positions. This process results in 210 (31,029 divided by 148) *max pooling features*, each one identifying the

21-bps sequence that obtained the highest value from the convolutional filter, in a specific 148-position interval of the original genome: the first max pooling feature will cover positions 1-148, the second will cover position 149-296, and so on. We graph the whole set of max pooling features for the complete data, Fig. 17.

Sequence 0 - 2,205

Sequence 2,205 - 2,210

Figure 17: cDNA visualization for the selected 210 21-bps-long sequences selected from the input dataset. Each sample is represented by a horizontal line of pixels. Colored pixels represent bases: G=green, C=blue, A=red, T=orange, missing=black. We divide the whole information, for visualization purposes; from visual inspection we can see the similarity of the patterns between the classes.

Analyzing the different sequence values appearing in the max pooling feature space, a total of 3,827 unique 21-bps cDNA sequences, that can potentially be very informative for identifying different virus strains. For example, sequence "AGGTAACAAACCAACCAACTT" is only found inside the class of SARS-CoV-2, in 59 out of 66 available samples. Sequence "CACGAGTAACTCGTC-TATCTT" is present only in SARS-CoV-2, in 63 out of the 66 samples.

The combination of the convolutional and max pooling layer allows the CNN to identify sequences even if they are slightly displaced in the genome (by up

to 148 positions). As some samples might present sequences that are displaced even more, in the next experiments we decided to just consider the relative frequency of the 21-pbs sequences identified at the previous step, creating a *sequence feature space*, to verify whether the appearance of specific sequences could be enough to differentiate between virus strains.

4.1. Example 1

We downloaded the dataset from the NGDC repository [6] on March 15[15] 2020. We removed repeated sequences and applied the whole procedure to translate the data into the sequence feature space. This leave us with a frequency table of 3,827 features with 583 samples (Table 3). Next, we ran a state-of-the-art feature selection algorithm [36], to reduce the sequences needed to identify different virus strain to the bare minimum. Remarkably, we are then able to classify exactly all samples using only 53 of the original 3,827 sequences, obtaining a 100% accuracy in a 10-fold cross-validation with a simpler and more traditional classifier, such as Logistic Regression.

Table 3: Organism, assigned label, and number of samples in the unique sequences obtained from the repository [6]. We use the NCBI organism naming convention [30].

Organism	Label	Number of Samples
SARS-CoV-2	0	96
MERS-CoV	1	236
HCoV-OC43	2	136
HCoV-229E	2	22
HCoV-EMC	2	6
HCoV-4408	2	2
HCoV-NL63	3	58
HCoV-HKU1	3	17
SARS-CoV	4	7
SARS-CoV P2	4	1
SARS-CoV HKU-39849	4	1
SARS-CoV GDH-BJH01	4	1

4.2. Example 2

We downloaded data from NCBI [22] on March 15$^{\text{th}}$ 2020, with the following query=*gene="ORF1ab" AND host="homo sapiens" AND "complete genome"*.

The query resulted in 407 non-repeated sequences (Table 4), with 68 sequences belonging to SARS-CoV-2. Then, we applied the whole procedure to translate the data into the sequence feature space, and we run the same state-of-the-art feature selection algorithm [36]. This give us a list of 10 different sequences: just checking for their presence is enough to differentiate between SARS-CoV-2 and other viruses in the dataset with a 100% accuracy. Each of the sequences only appears in SARS-CoV-2.

Table 4: Organism, assigned label, and number of samples in the unique sequences obtained from the repository NCBI [30].

Virus	Label	Number of Samples
SARS-CoV-2	0	68
MERS-CoV	1	180
HCoV-OC43	1	105
HCoV-NL63	1	29
HCoV-HKU1	1	13
HCoV-4408	1	2
HCoV-229E	1	3
HCoV-EMC	1	3
HAstV-VA1	1	1
HAstV-BF34	1	1
HMO-A	1	1
HAstV-SG	1	1

4.3. Example 3

We downloaded data from NCBI [22] on March 17$^{\text{th}}$ 2020, with the following query=*"virus AND host="homo sapiens" AND "complete genome"*, restricting the size from 1,000 to 35,000. This gives us a total of 20,603 results, where only 32 samples are SARS-CoV-2 samples and 20,571 are from other taxa, including; Hepatitis B, Dengue, Human immunodeficiency, Human orthopneumovirus, Enterovirus A, Hepacivirus C, Chikungunya virus, Zaire ebolavirus, Human respirovirus 3, Orthohepevirus A, Norovirus GII, Hepatitis delta virus, Mumps rubulavirus, Enterovirus D, Zika virus, Measles morbillivirus, Enterovirus C,

Table 5: Sequences that only exist in SARS-CoV-2, that help differentiate between the virus and other taxa as displayed in Table 4.

TAGCACTCTCCAAGGGTGTTC
CATCTACTGATTGGACTAGCT
AATGAATTATCAAGTTAATGG
CACGTAGGAATGTGGCAACTT
TGAGCAGTGCTGACTCAACTC
CAACTTTTAACGTACCAATGG
CTAAAGCATACAATGTAACAC
GATGGTCAAGTAGACTTATTT
TGCCACTTGGCTATGTAACAC
TATTAGTGATATGTACGACCC

Human T-cell leukemia virus type I, Yellow fever virus, Adeno-associated virus, rhinovirus (A, B and C), for more than 900 viruses. Then, we we applied the whole procedure to translate the data into the sequence feature space and run the feature reduction algorithm [36]. This results in 2 sequences of 21 bps: just by checking for their presence, we are able to separate SARS-CoV-2 from the rest of the samples with a 100% accuracy. The sequences are: **AATAGAAGAATTATTCTATTC** and **CGATAACAACTTCTGTGGCCC**.

4.4. Example 4

From the GISAID repository [37], we downloaded the last 323 sequences available for SARS-CoV-2, from different countries. Then, we calculate the frequency table of the 21-bps sequences from examples 2 and 3, to see which sequences remain and could be use for detection. The results are in Table 6 in percentage of appearance in the GISAID sequences.

Table 6: Frequency table in percentages for the sequences in examples 2 and 3 in the 323 sequences from GISAID [37].

TAGCACTCTCCAAGGGTGTTC	100.00%
AATGAATTATCAAGTTAATGG	100.00%
TATTAGTGATATGTACGACCC	100.00%
AATAGAAGAATTATTCTATTC	100.00%
CACGTAGGAATGTGGCAACTT	99.69%
CAACTTTTAACGTACCAATGG	99.69%
CTAAAGCATACAATGTAACAC	99.69%
GATGGTCAAGTAGACTTATTT	99.69%
TGAGCAGTGCTGACTCAACTC	99.38%
CGATAACAACTTCTGTGGCCC	99.07%
CATCTACTGATTGGACTAGCT	98.76%
TGCCACTTGGCTATGTAACAC	95.04%

4.5. Biological features and molecular techniques

After deep learning analysis, we identify that the sequence TAGCACTCTCCAAGGGTGTTC is exclusive for SARS-CoV-2 and shows a frequency of 100% in viral genomes available from different countries in the GISAID [37] and NCBI [22]. Using NC045512.2 as reference SARS-CoV-2 sequence, we identify this unique sequence is located from 25604 to 25624 nucleotides in ORF3a gene. In SARS-CoV, this gene encodes a protein of 274 aa, that is related with necrotic cell death [38], chemokine production, inflammatory response [39] and may play an important role in virus life cycle [40]. With this information, we design a specific primer set for detection of SARS-CoV-2 using Primer3plus [41]. We use TAGCACTCTCCAAGGGTGTTC as forward primer and GCAAAGCCAAAGCCTCATTA as reverse primer. Then, we run an *In silico PCR* test using FastPCR 6.7 [42] with default parameters, this yields the results from Fig. 18.

These primers (Forward 5' TAGCACTCTCCAAGGGTGTTC 3' and Reverse 5' GCAAAGCCAAAGCCTCATTA 3') could identify and differentiate SARS-CoV-2 from other coronavirus species through the PCR method. Furthermore, we propose to create a multiplex PCR using the 21 nt. unique sequences for SARS-CoV-2 identified through deep learning to develop accurate molecular diagnostic techniques. However, it is necessary to test it in laboratory and carry out the validation with patients samples.

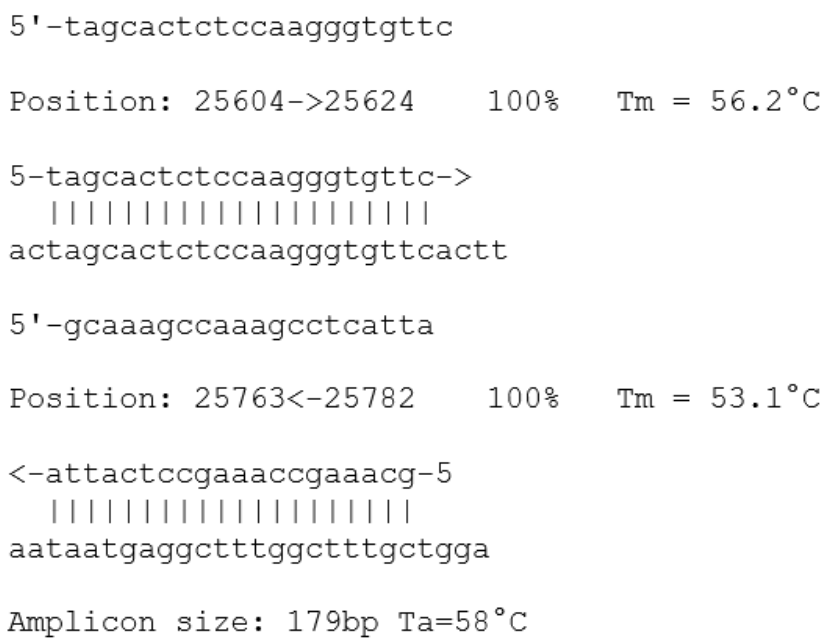


Figure 18: In silico PCR Test results using TAGCACTCTCCAAGGGTGTTC and GCAAAGCCAAAGCCTCATTA sequences as primers in NC045512.2 as reference SARS-CoV-2 sequence using FastPCR 6.7 program [42].

Severity Identification

Experiment 5: Severity Detection

The data collected from the GISAID repository [37], also reports metadata, including the status of patients. While most of the metadata is missing, we selected 169 patients for which the status is reported; in the dataset, 52 are annotated as *asymptomatic* and 117 as *hospitalized*. We reached to the submitters of some of the sequences, and they reported that *hospitalized* meant that the patients presented evident symptoms of SARS-CoV-2, and could thus be considered *symptomatic*. We then applied the previously described methodology to discover specific sequences to separate asymptomatic from symptomatic (hospitalized) patients, and reduced the number of 21-bps sequences to the necessary minimum, using a feature reduction algorithm [36]. The algorithm ultimately returns an optimal set of 32 sequences, of 21 bps each: Simply checking for their presence inside a patient sample makes it possible to separate asymptomatic from symptomatic patients with 94% accuracy.

For each of the discovered sequences we then calculate the frequency of appearance in both asymptomatic and symptomatic patients. The results are reported in Table 7. It is important to notice that the sequences discovered for this experiment do not overlap with those identified in previous validations, as the objective of this last test is considerably different: While in previous experiments the aim was to separate SARS-CoV-2 samples from other virus strains, here the goal is to separate SARS-CoV-2 patients that require hospitalization from those who do not.

In contrast to Experiments 1-4 where the 21-length bps sequences were scattered, in Experiment 5, seven of the sequences are clustered together in the symptomatic cases Fig. 19, and four of the sequences in asymptomatic cases Fig. 20. These sequences are located in the same region of ORF1ab gene, but contain a mutation that discriminates between symptomatic and asympotomatic patients ($c.11083G > T$). This transversion results in the substitution of leucine to phenylalanine (p.L3606F). A previous report identified the same mutation in

Table 7: Appearance frequency table for the 169 sequences. We only show the sequences with the biggest percentage differences between symptomatic and asymptomatic patients.

Sequence	Asymptomatic	Symptomatic	Absolute Difference
TTTTTATGAAAATGCCTTTTT	94.23%	17.95%	76.28%
TTTATGAAAATGCCTTTTTAC	94.23%	17.95%	76.28%
TTTTATGAAAATGCCTTTTTA	94.23%	17.95%	76.28%
TTTTTTTTTTATGAAAATGCC	94.23%	17.95%	76.28%
TGTATGAAAATGCCTTTTTAC	5.77%	81.20%	75.43%
GTATGAAAATGCCTTTTTACC	5.77%	81.20%	75.43%
TTTTGTATGAAAATGCCTTTT	5.77%	81.20%	75.43%
GTTCTTTTTTTTGTATGAAAA	5.77%	81.20%	75.43%
TTTGTTCTTTTTTTTGTATGA	5.77%	81.20%	75.43%
TTGTATGAAAATGCCTTTTTA	5.77%	81.20%	75.43%
TGTTCTTTTTTTTGTATGAAA	5.77%	81.20%	75.43%
AAACCAACCAACTTTCGATCT	19.23%	63.25%	44.02%
TTAAAGGTTTATACCTTCCCA	0.00%	26.50%	26.50%
TCGTAACTATATAGCACAAGT	26.92%	0.85%	26.07%
GATCTGTTCTCTAAACGAACT	76.92%	94.87%	17.95%
CAACCAACTTTCGATCTCTTG	59.62%	69.23%	9.62%

the ORF1ab gene [43]. ORF1ab proteins play an important role in pathogenesis and viral replication. These might be through the interaction of structural and non-structural proteins, besides the regulatory sequences in viral RNA. Furthermore, it has been described that mutations in ORF1ab are positively selected during trans-species transmission of SARS-CoV and SARS-like coronaviruses. Thus, we suggest that the missense mutation in ORF1ab gene ($c.11083G > T$) could alters the viral load during the infection between symptomatic and asympotomatic patients [44].

Fig. 21. In summary, we get the Table 8. With one extra symptomatic case the presents a missing value, giving the following sequence; **TTTGTTCTTTTTTT-TNTATGAAAATGCCTTTTTACC**.

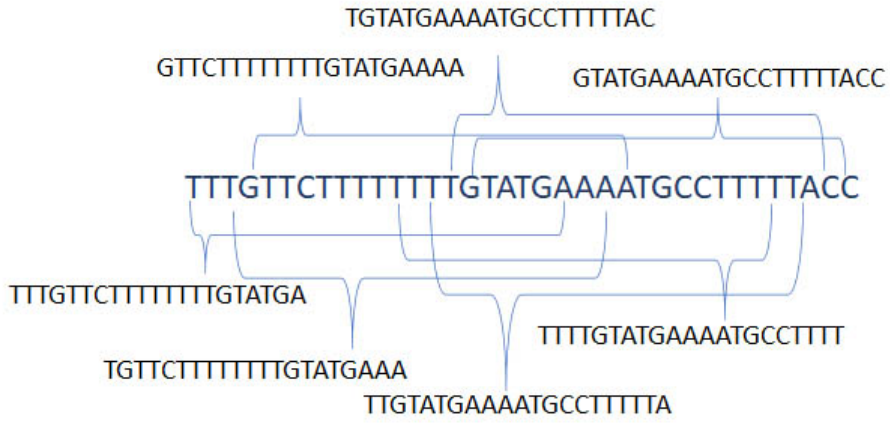


Figure 19: 7 of the first 16 sequences clustered together in a 36-bps sequence, which primarily appears in symptomatic cases.

Table 8: Frequency of appearance of the sequences in asymptomatic and symptomatic cases.

Sequences	Asymptomatic	Symptomatic
TTTTTTTTTTATGAAAATGCCTTTTTAC	94.23%	5.77%
TTTTTTTTGTATGAAAATGCCTTTTTAC	17.95%	81.20%

Finally, using NC045512.2 as the reference SARS-CoV-2 sequence, and **TTTTTTTT[G/T]TATGAAAATGCCTTTTTAC** as target sequence,

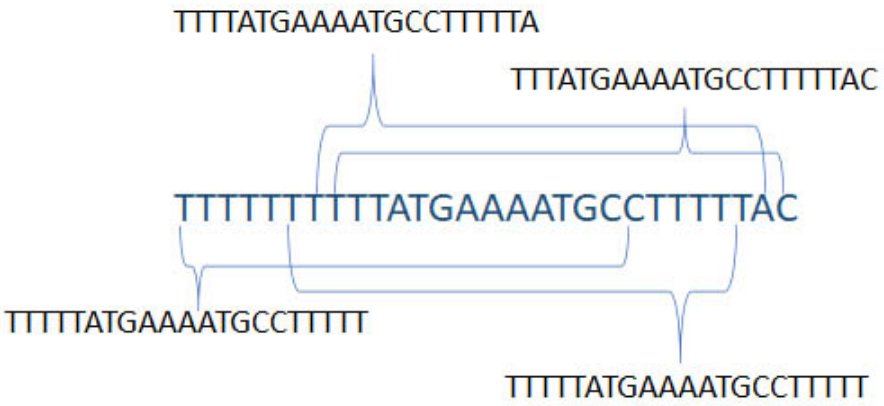


Figure 20: 4 of the first 16 sequences clustered together in a 28-bps sequence, which primarily appears in asymptomatic cases.

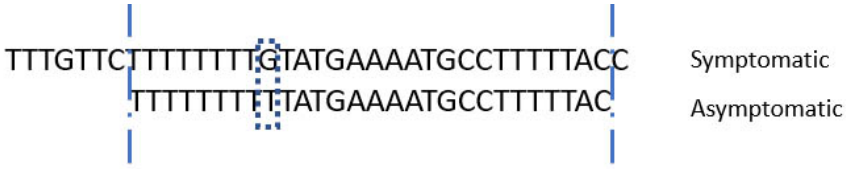


Figure 21: Single-nucleotide polymorphism (SNP) between two discovered sequences, that separates asymptomatic from symptomatic patients with a 85% accuracy.

we generate a primer set using *Primer3plus* [41]. This outputs **5'TTCCAAAGTGCAGTGAAAAGAA3'** as forward primer and **5'TTGCAAAAGCAGACATAGCAA3'** as reverse primer with a total length of 175 bps. Then, we run an *in-silico PCR* test using FastPCR 6.7 [42] with default parameters, this yields the results reported in Fig. 22. Nevertheless, the results of the PCR Amplicons will need to be sequenced, to differentiate between the two possible sequences.

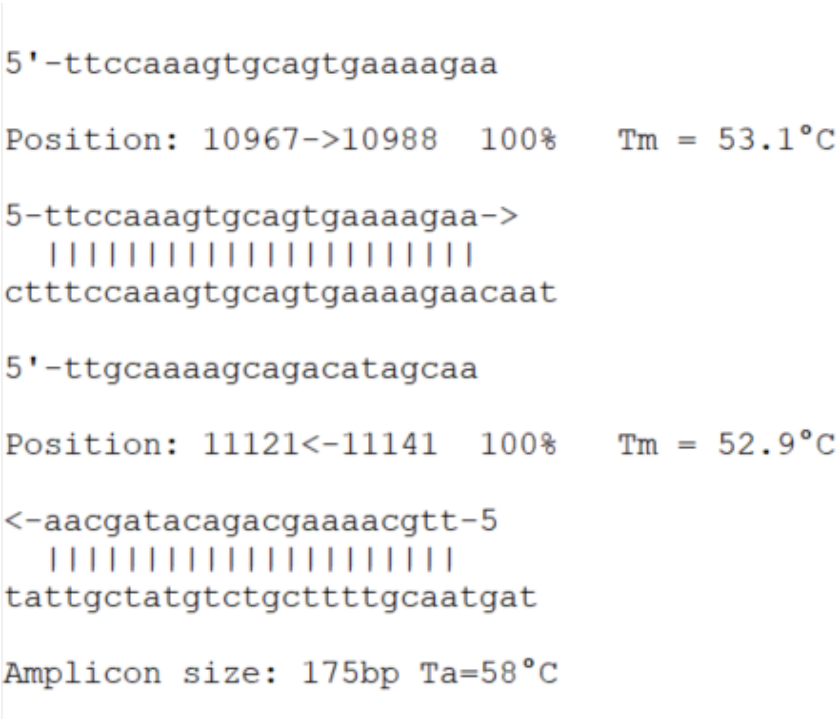


Figure 22: In-silico PCR test results from the FastPCR 6.7 software [42] using sequences TTCCAAAGTGCAGTGAAAAGAA and TTGCAAAAGCAGACATAGCAA as primers, in NC045512.2 used as a reference SARS-CoV-2 sequence.

5. Conclusion

Being able to reliably identify SARS-CoV-2 and distinguish it from other similar pathogens is important to contain its spread. The time of processing samples and the availability of reliable diagnostic tests is a challenge during an outbreak. Developing innovative diagnostic tools that target the genome to improve the identification of pathogens, can help reduce health costs and time to identify the infection, instead of using unsuitable treatments or testing. Moreover, it is necessary to perform an accurate classification to identify the different species of Coronavirus, the genetic variants that could appear in the future, and the co-infections with other pathogens.

Given the high transmissibility of the SARS-CoV-2, the proper diagnosis of the disease is urgent, to stop the virus from spreading further. Considering the false negatives given by the standard nucleic acid detection, better implementations such as using deep learning are necessary in order to to properly detect the virus. While the accuracy of current nucleic acid testing is around 30-50%, and CT scans with deep learning go up at 83%, we believe that the use of the sequences detected by a CNN-based system has the potential to improve the accuracy of the diagnosis above 95%.

Our results, show that by targeting only 12 21-bps specific sequences, we are able to distinguish SARS-CoV-2, from any other virus ($> 99\%$). In addition, with 85% accuracy is possible to predict if an infected person will need to be hospitalized or will be asymptomatic. These findings could help to identify patients with SARS-CoV-2 that are susceptible to develop severe acute respiratory infection and make a better clinical management. Nevertheless, our conclusions hold only for the data currently at our disposal. Further testing is necessary to confirm these promising results so it is essential to create multidisciplinary groups that work to stop the outbreak. Finally, as an interesting remark, by comparing the discovered sequences against other hosts, we noticed that from the 12 sequences exclusive to SARS-CoV-2, 1 of them appears in all of the 9 sequences from *Manis Javanina*. In contrast, 5 of the sequences of SARS-CoV-2

355 appear in the only sample available from *Rhinolophus Affinis*. The addition of these sequences sum up to 6 of the 12 sequences that we used to characterized the SARS-CoV-2. This is consistent with the findings of Zhang et al. [45], and could point to the zootonic origin of the virus.

References

[1] P. C. Woo, Y. Huang, S. K. Lau, K.-Y. Yuen, Coronavirus genomics and bioinformatics analysis, viruses 2 (8) (2010) 1804–1820.

[2] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, et al., Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, The Lancet 395 (10224) (2020) 565–574.

[3] V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, et al., Detection of 2019 novel coronavirus (2019-ncov) by real-time rt-pcr, Eurosurveillance 25 (3) (2020).

[4] D. K. Chu, Y. Pan, S. Cheng, K. P. Hui, P. Krishnan, Y. Liu, D. Y. Ng, C. K. Wan, P. Yang, Q. Wang, et al., Molecular diagnosis of a novel coronavirus (2019-ncov) causing an outbreak of pneumonia, Clinical chemistry (2020).

[5] D. A. Marston, L. M. McElhinney, R. J. Ellis, D. L. Horton, E. L. Wise, S. L. Leech, D. David, X. de Lamballerie, A. R. Fooks, Next generation sequencing of viral rna genomes, BMC genomics 14 (1) (2013) 444.

[6] Beijing Institute of Genomics, Chinese Academy of Science, China National Center for Bioinformation & National Genomics Data Center, `https://bigd.big.ac.cn/ncov/?lang=en`, online; accessed 27 January 2020 (2013).

[7] W. H. Organization, WHO report Coronavirus disease 2019 (COVID-19), World Health Organization,, Geneva :, 2020., licence : CC BY-NC-SA 3.0 IGO.

[8] Y. Wang, H. Kang, X. Liu, Z. Tong, Combination of rt-qpcr testing and clinical features for diagnosis of covid-19 facilitates management of sars-cov-2 outbreak, Journal of Medical Virology (2020).

[9] H. C. Metsky, C. A. Freije, T.-S. F. Kosoko-Thoroddsen, P. C. Sabeti, C. Myhrvold, Crispr-based surveillance for covid-19 using genomically-comprehensive machine learning design, bioRxiv (2020).

[10] M. Wang, Q. Wu, W. Xu, B. Qiao, J. Wang, H. Zheng, S. Jiang, J. Mei, Z. Wu, Y. Deng, et al., Clinical diagnosis of 8274 samples with 2019-novel coronavirus in wuhan, medRxiv (2020).

[11] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, et al., A deep learning algorithm using ct images to screen for corona virus disease (covid-19), medRxiv (2020).

[12] J. Y. Kim, P. G. Choe, Y. Oh, K. J. Oh, J. Kim, S. J. Park, J. H. Park, H. K. Na, M.-d. Oh, The first case of 2019 novel coronavirus pneumonia imported into korea from wuhan, china: implication for infection prevention and control measures, Journal of Korean Medical Science 35 (5) (2020).

[13] W. R. Pearson, [5] rapid and sensitive sequence comparison with fastp and fasta (1990).

[14] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool, Journal of molecular biology 215 (3) (1990) 403–410.

[15] L. Pinello, G. Lo Bosco, G.-C. Yuan, Applications of alignment-free methods in epigenomics, Briefings in Bioinformatics 15 (3) (2014) 419–430.

[16] S. Vinga, J. Almeida, Alignment-free sequence comparison—a review, Bioinformatics 19 (4) (2003) 513–523.

[17] D. Bzhalava, J. Ekström, F. Lysholm, E. Hultin, H. Faust, B. Persson, M. Lehtinen, E.-M. de Villiers, J. Dillner, Phylogenetically diverse tt virus viremia among pregnant women, Virology 432 (2) (2012) 427–434.

[18] N. G. Nguyen, V. A. Tran, D. L. Ngo, D. Phan, F. R. Lumbanraja, M. R. Faisal, B. Abapihi, M. Kubo, K. Satou, et al., Dna sequence classification by convolutional neural network, Journal of Biomedical Science and Engineering 9 (05) (2016) 280.

[19] R. Rizzo, A. Fiannaca, M. La Rosa, A. Urso, A deep learning approach to dna sequence classification, in: International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, Springer, 2015, pp. 129–140.

[20] A. Tampuu, Z. Bzhalava, J. Dillner, R. Vicente, Viraminer: Deep learning on raw dna sequences for identifying viral genomes in human samples, PloS one 14 (9) (2019).

[21] J. Ren, K. Song, C. Deng, N. A. Ahlgren, J. A. Fuhrman, Y. Li, X. Xie, F. Sun, Identifying viruses from metagenomic data by deep learning, arXiv preprint arXiv:1806.07810 (2018).

[22] S. T. Sherry, M.-H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, K. Sirotkin, dbsnp: the ncbi database of genetic variation, Nucleic acids research 29 (1) (2001) 308–311.

[23] C. d. S. Ribeiro, M. Y. van Roode, G. B. Haringhuizen, M. P. Koopmans, E. Claassen, L. H. van de Burgwal, How ownership rights over microorganisms affect infectious disease control and innovation: a root-cause analysis of barriers to data sharing as experienced by key stakeholders, PloS one 13 (5) (2018).

[24] J. H. Simon, E. Claassen, C. E. Correa, A. D. Osterhaus, Managing severe acute respiratory syndrome (sars) intellectual property rights: the possible role of patent pooling, Bulletin of the World Health Organization 83 (2005) 707–710.

[25] C. d. S. Ribeiro, M. P. Koopmans, G. B. Haringhuizen, Threats to timely sharing of pathogen sequence data, Science 362 (6413) (2018) 404–406.

[26] A. Lopez-Rincon, A. Tonda, M. Elati, O. Schwander, B. Piwowarski, P. Gallinari, Evolutionary optimization of convolutional neural networks for cancer mirna biomarkers classification, Applied Soft Computing 65 (2018) 91–100.

[27] A. Vabret, T. Mourez, S. Gouarin, J. Petitjean, F. Freymuth, An outbreak of coronavirus oc43 respiratory infection in normandy, france, Clinical infectious diseases 36 (8) (2003) 985–989.

[28] L.-J. Cui, C. Zhang, T. Zhang, R.-J. Lu, Z.-D. Xie, L.-L. Zhang, C.-Y. Liu, W.-M. Zhou, L. Ruan, X.-J. Ma, et al., Human coronaviruses hcov-nl63 and hcov-hku1 in hospitalized children with acute respiratory infections in beijing, china, Advances in virology 2011 (2011).

[29] F. Zeng, C. Chan, M. Chan, J. Chen, K. Chow, C. Hon, K. Hui, J. Li, V. Li, C. Wang, et al., The complete genome sequence of severe acute respiratory syndrome coronavirus strain hku-39849 (hk-39), Experimental Biology and Medicine 228 (7) (2003) 866–873.

[30] I. Mizrachi, Genbank: the nucleotide sequence database, The NCBI Handbook [Internet], updated 22 (2007).

[31] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[32] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org (2015).
URL `https://www.tensorflow.org/`

[33] F. Chollet, et al., Keras, `https://keras.io` (2015).

[34] A.-M. Šimundić, Measures of diagnostic accuracy: basic definitions, Ejifcc 19 (4) (2009) 203.

[35] J. N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, Journal of Thoracic Oncology 5 (9) (2010) 1315–1316.

[36] A. Lopez-Rincon, M. Martinez-Archundia, G. U. Martinez-Ruiz, A. Schoenhuth, A. Tonda, Automatic discovery of 100-mirna signature for cancer classification using ensemble feature selection, BMC bioinformatics 20 (1) (2019) 480.

[37] Y. Shu, J. McCauley, Gisaid: Global initiative on sharing all influenza data–from vision to reality, Eurosurveillance 22 (13) (2017).

[38] C.-S. Shi, N. R. Nabar, N.-N. Huang, J. H. Kehrl, Sars-coronavirus open reading frame-8b triggers intracellular stress pathways and activates nlrp3 inflammasomes, Cell death discovery 5 (1) (2019) 1–12.

[39] N. Kanzawa, K. Nishigaki, T. Hayashi, Y. Ishii, S. Furukawa, A. Niiro, F. Yasui, M. Kohara, K. Morita, K. Matsushima, et al., Augmentation of chemokine production by severe acute respiratory syndrome coronavirus 3a/x1 and 7a/x4 proteins through nf-κb activation, FEBS letters 580 (30) (2006) 6807–6812.

[40] K. Padhan, C. Tanwar, A. Hussain, P. Y. Hui, M. Y. Lee, C. Y. Cheung, J. S. M. Peiris, S. Jameel, Severe acute respiratory syndrome coronavirus orf3a protein interacts with caveolin, Journal of General Virology 88 (11) (2007) 3067–3077.

[41] A. Untergasser, H. Nijveen, X. Rao, T. Bisseling, R. Geurts, J. A. Leunissen, Primer3plus, an enhanced web interface to primer3, Nucleic acids research 35 (suppl_2) (2007) W71–W74.

[42] R. Kalendar, D. Lee, A. H. Schulman, et al., Fastpcr software for pcr primer and probe design and repeat search, Genes, Genomes and Genomics 3 (1) (2009) 1–14.

[43] T. Phan, Genetic diversity and evolution of sars-cov-2, Infection, Genetics and Evolution 81 (2020) 104260.

[44] R. L. Graham, J. S. Sparks, L. D. Eckerle, A. C. Sims, M. R. Denison, Sars coronavirus replicase proteins in pathogenesis, Virus research 133 (1) (2008) 88–100.

[45] Y.-Z. Zhang, E. C. Holmes, A genomic perspective on the origin and emergence of sars-cov-2, Cell (2020).