# Human DNA AI Model to Predict COVID-19 Symptomatic or **Asymptomatic Percentages**

Peter Savier Oropeza-Martínez<sup>1</sup>, Luis Manuel Gaggero-Sager<sup>1\*</sup>, Haydeé Rosas-Vargas<sup>2</sup>

<sup>1</sup>Center of Research in Engineering and Applied Sciences, the Autonomous University of the State of Morelos, Mexico <sup>2</sup>Medical Research Unit of Human Genetics, Pedriatic Hospital, National Medical Center "Siglo XXI", IMSS, Mexico CIty, Mexico

Article History:	Submitted: 16.07.2021	Accepted: 30.07.2021	Published: 06.08.2021
ABSTRACT The current paper proposes to use convolutional neu-		80.986% and 19.014% respectively and the model was validated through an online survey of individuals, giving	

ral networks (CNN) to analyze human genome Single Nucleotide Variants (SNVs) from nuclear Deoxyribonucleic Acid (DNA) and Mitochondrial Deoxyribonucleic Acid (mtDNA) presented as a 2D image structure to understand if the answer to COVID-19 severities can be found in the human genome. That methodology was implemented with 447 Mexican population samples. From results two main groups were formed divided in symptomatic and asymptomatic cases composed by

#### **INTRODUCTION**

In December 2019, there was an outbreak of pneumonia of unknown cause in Wuhan, Hubei province, China. This disease outbreak attacked locals with complications as: Fever, malaise, dry cough, shortness of breath and respiratory failure (CDC Weekly, 2020). On March 11, the disease was already in more than 100 territories worldwide, and it was recognized as a pandemic by the World Health Organization (WHO, 2020). The number of confirmed cases continued to grow worldwide. To prevent the spread of the virus, governments have imposed travel restrictions, quarantines, lockdowns, social isolation, cancellation of events, and closure of establishments. The pandemic is having a disruptive socioeconomic effect (Konrad N, 2020).

There have been different investigations about the causes of the COVID-19 severity in the population related to research in blood type, previous or current smoking, among others (Al-Khikani FHO, 2020, Constantine V and Nikitara K, 2020, Khan AH, et al., 2020) and there are still no convincing results about the causes.

Genomics is a wide research field in numerous diseases and COVID-19 is not an exception. Found significant differences in the structural genomics of patients who had severe reactions to the virus vs. general population (Toh C and Brody JP, 2020). However, the test developed does not have a predictive value, so further research is still needed. An alternative approach is exemplified in the work of 8 who, based on a genome wide association study (GWAS) in which mortality was taken as the primary endpoint in patients with COVID-19, defined (Hu J, et al., 2021) super variants that reflect the interaction of multiple loci associated with an increased risk of mortality. Another important finding is described (Scozzi D, et al., 2021), who evaluated 97 patients with COVID-19 at Barnes-Jewish hospital by measuring their circulating mtDNA levels on the first day of their hospital stay. They found that mtDNA levels were much higher in patients who eventually died or were admitted to the intensive care unit. This association held independently of the patient's age, sex, and underlying health conditions.

Genome genotyping is an open door to long periods of research of certain diseases or specific conditions for prognostic tests based a 91.89% of accuracy.

Keywords: Human genome, DNA, Artificial Intelligence (AI)

Correspondence: Luis Manuel Gaggero-Sager, Center of Research in Engineering and Applied Sciences, the Autonomous University of the State of Morelos, Mexico. E-mail: Igaggero@uaem.mx

on people's DNA to give information about family history, ancestry, personal identity, and health info. Companies like SIKUENS Genetics, 23andMe, CRI Genetics, Ancestry DNA, among others, offer this type of services and they constantly move forward with the investigation of a growing number of Single Nucleotide Variants (SNVs) related to certain diseases or conditions to offer more information panels for their reports (23andMe, 2016). Similar to these companies, biomedical and biotechnological research institutes worldwide need long periods of research to determine the entire genome related to a certain disease. This is the reason why we moved forward in a computer science technique where it can provide the pattern recognition of a large panel of genomic SNVs through 2D image structures containing information and without the need of large periods of time and hardware resources consumption.

There are different 2D data containing structures, such as Data Matrix with a capacity of 3116 numeric characters, Maxi Code with a capacity of 93 ASCII characters, PDF 417 with 2725 numeric characters, QR Code with 7089 numeric characters, Codablock F with 5450 numeric characters, Aztec Code with 3832 numeric characters, as well as many other structures within this range of data capacity (Upkeep, 2019).

In different research, data capacity in 2D structures has been improved using different techniques for different types of applications, such as DNA QR coding for security systems (Singh SP and Naidu ME, 2020), DNA species identification (Naulia T, 2015; Gogoi B, et al., 2020), but data capacity in 2D structures is still a challenge for DNA information encompassing 642,824 SNV's with two alleles per each SNV for a total amount of data of 1,285,648.

#### MATERIALS AND METHODS

Based on a previous work (Quiceno H, 2019) we considered using a similar technique for a SNVs omic analysis. The proposed structure is an image of 802 pixels of width and 802 pixels of height for a total amount of data of 643,204 pixels. The genome was represented in each pixel as a combination of the two alleles per SNV and can be visualized (Figure 1).



# Figure 1: New DNA human genome representation of a Mexican sample

#### Training stage

To the unsupervised CNN it is provided a set of DNA of individuals in the Mexico population that can be considered as having contracted COVID-19 together with the information if the individual has developed symptoms of contracting COVID-19.

The DNA dataset is important to mention that it is extracted from the oral epithelium through cells from the walls of the mouth, not from saliva using a scraping methodology. Such sample is genotyped and can be replicated through a list of SNV's published in GitHub, 2021 of the human DNA including nuclear and mitochondrial DNA.

Once all the datasets were transformed in their new image representation, clusters of images were generated through the unsupervised learning algorithm k-means. The pseudocode and equations who rule the algorithm can be consulted in Li Y and Wu H, 2012.

Once the clusters were formed, the CNN learned the patterns in those clusters and ended correctly classifying the validation dataset to the corresponding clusters. The CNN architecture proposed for the current research was implemented in Python using Google's Tensorflow and Keras for the supervised learning and can be visualized in *Figure 2*. The learning process of the CNN algorithm can be consulted in Li Y and Wu H, 2012.



#### Figure 2: Proposed CNN general architecture

When analyzing the dataset conformed by 447 images, in first instance, we found a 76.95% of images in a main cluster of related ones and another main cluster with 23.04% of the images; this relations between cluster are based on a defined threshold greater than 95% of probability ( $\mathbf{Q} \ge 0.95$ ) (Yamashita R, *et al.*, 2018).

In Mexico, the tests have only been used as a diagnostic method and not as a tool to predict the severity of the response to the disease, in addition to the percentage of false negatives and false positives, therefore, we considered that only using the results of a positive test of COVID would not be reliable as a criterium to decide if an individual has contracted COVID-19. To overcome this difficulty, we relied on interviews where the respondents provide the information if they have been tested and their tests turned positive, or if they have developed symptoms of COVID or if they have been in close contact with family members that have contracted COVID.

#### Validation stage

In a second stage, it was provided to the CNN a set of DNA also in the Mexican population, without any information if the individual has contracted COVID-19 and the CNN must identify which ones would develop symptoms. Two clusters of asymptomatic and symptomatic individuals were obtained composed of 80.986% and 19.014% of images.

## **RESULTS AND DISCUSSION**

After the two main clusters were formed, a survey was designed and sent to the individuals, clients of the company, but only 37 answers were received. The survey was composed by one question with 5 options, and they were:

- 1. I have not suffered from COVID-19.
- 2. I have lived with infected people and I have not contracted COVID-19.

3. I have been infected with COVID-19 and I have not presented symptoms.

4. I have been infected with COVID-19 and I have had mild symptoms.

5. I have been infected with COVID-19 and have had severe symptoms.

Once the answers were received, we divided the answers into 3 groups. The groups are:

• Answer 1 is Uncertain, because when the moment of the survey the person can still not retrieve de symptoms due to has not been infected yet or otherwise will not retrieve symptoms.

• Answers 2 and 3 are Asymptomatic.

• Answers 4 and 5 are Symptomatic.

Finally, when the answers of the survey were categorized in previous mentioned groups and the CNN model learned the clusters formed by k-means, the DNA of the people who answered the survey were predicted by the CNN model and the results are 25 precise predictions of Asymptomatic people, 9 precise predictions of Symptomatic people and 3 ambiguous predictions of Symptomatic people who answered in the survey the Answer 2, "I have lived with infected people and I have not contracted COVID-19" or in other words asymptomatic. As well, if the model predicted an asymptomatic or symptomatic case and compared to the individual answer was an uncertain class, the result was considered as good prediction.

#### CONCLUSION

It can be concluded that, observing the results presented in this research, the clusters formed are an exceptionally good approximation to COVID-19 statistics of virus severities in Mexican population, but we cannot conclude the clusters are related to COVID-19 without a validation process. So, the model was validated through the individual's survey and resulted with a 91.89% of accuracy based on those answers considering that that the answers are not a clinical test result as Polymerase Chain Reaction (PCR). This is the reason why we can interpret the 3 wrong predictions as ambiguous because the person can be pre-symptomatic and has still not developed symptoms yet. In this manner, the accuracy of the CNN model would be much higher but for the mean time we can conclude that the clusters formed are related with COVID-19 due to the validation process; as well we can conclude that the causes of COVID-19 severities lie on human genetics so it may or may not potentially allow the virus to advance.

Finally, considering future work, it will be incorporating datasets from different populations around the world and observe the clusters and percentages formed with help of data access of the human genome as well as clinical PCR tests or surveys of the population corresponding to those DNA data for model validation.

### ACKNOWLEDGMENT

We want to thank SIKUENS Genetics (Grupo SIKUENS S.A. de C.V.) as the provider of the anonymous Mexican genetic datasets employed in this research and it's CEO Eduardo Tena Alavez for opening gates to this great finding. As well, we want to thank Prof. Enrique Pujals for contributing for the discussion and structure of this paper.

#### REFERENCES

- The 2019-nCoV Outbreak Joint Field Epidemiology Investigation Team, Qun Li. An Outbreak of NCIP (2019-nCoV) Infection in China-Wuhan, Hubei Province, 2019-2020. China CDC Weekly. 2020; 2(5): 79-80.
- 2. Coronavirus Disease 2019 (COVID-19) Situation Report-51. WHO. 2020.
- 3. Konrad N. Here comes the coronavirus pandemic. The New York Times. 2020.
- 4. Al-Khikani FHO. The role of blood group in COVID-19 infection: More information is needed. J Nat and Sci Med. 2020; 3(3): 225-226.
- 5. Constantine V, Nikitara K. COVID-19 and smoking: A systematic review of the evidence. Tob Induc Dis. 2020; 18: 20.
- Khan AH, Sultana S, Hossain S, Hasan MT, Ahmed HU, Sikder T. The impact of COVID-19 pandemic on mental health and wellbeing among home-quarantined Bangladeshi students: A cross-sectional pilot study. J Affect Disord. 2020; 277(1): 121-128.
- Toh C, Brody JP. Evaluation of a genetic risk score for severity of COVID-19 using human chromosomal-scale length variation. Hum Genomics. 2020; 14: 36.
- 8. Hu J, Li C, Wang S, Lis T, Zhang H. Genetic variants are identified to increase risk of COVID-19 related mortality from UK Biobank data. Hum Genomics. 2021; 15: 10.

- 9. Scozzi D, Cano M, Ma L, Zhou D, Zhu JH, O'Halloran JA, *et al.* Circulating mitochondrial DNA is an early indicator of severe illness and mortality from COVID-19. JCI Insight. 2021. 6(4), e143299.
- 10. New genetic variants found to influence likelihood of being a morning person. 23andMe. 2016.
- 11. What is a 2D Barcode? UpKeep Maintainence Glossory. 2019.
- 12. Singh SP, Naidu ME. DNA QR coding for data security using DNA sequence. Int J Inf Tecnol. 2020; 12: 571-576.
- Naulia T. DNA QR code scanner for identifying the species origin of meat products. ISICO. 2015.
- 14. Gogoi B, Wann SB, Saikia SP. DNA barcodes for delineating Clerodendrum species of North East India. Sci Rep. 2020; 10(1): 13490.
- 15. Quiceno H. DNA--PNG. JamMarHer/DNA--PNG. 2019.
- 16. Oropeza-Martínez. SNP list in the Mexican population to create a 2D image structure. GitHub. 2021.
- 17. Li Y, Wu H. A clustering method based on k-means algorithm. Physics Procedia. 2012; 25: 1104-1109.
- Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: An overview and application in radiology. Insights Imaging, 2018; 9: 611-629.