PREDICTING COVID-19 TRAJECTORY USING MACHINE LEARNING

ZAINAB ABBAS ABDULHUSSEIN
ALWAELI
Department of Computer engineering
Altinbas univeristy
Istanbul, Turkey
zainb.abbas86@yahoo.com

Asst. prof. Dr. Abdullahi Abdu Ibrahim
Department of Computer engineering
Altinbas univeristy
Istanbul, Turkey
abdullahi.ibrahim@altinbas.edu.tr

***Abstract*— The pandemic caused by COVID-19 in 2020 triggered a devastating effect on the economy and health of the world population, whose social implications for the next few years are still uncertain. Two types of standard tests are used to detect COVID-19: the viral test that indicates whether the patient is infected and the antibody test that allows us to observe if the patient has previously had an infection. These tests employ techniques such as reverse transcription and polymerase chain reaction (RT-PCR), immunochromatographic lateral flow or rapid test, and ELISA-type immunoassay In this paper we have designed and implemented a system whose main purpose is to detect the rise of Covid-19 cases using disruptive technologies such as artificial intelligence and intelligent computing, manifested through machine learning (Machine Learning) and deep learning (Deep Learning). Combined with data science, Big Data and advanced data analytics, among others that present various research and development options, it can help the early detection of COVID-19 through the search for relevant characteristics that allow the scientific community identify biochemical, molecular and cellular factors that facilitate the early detection of the virus in its different states of infection, incubation, propagation and treatments to be used**

Keywords—component, formatting, style, styling, insert *(key words)*

I. INTRODUCTION

A complementary alternative to the aforementioned techniques is the use of artificial intelligence (AI), Big Data and other disruptive technologies related to the analysis of massive data, which allow detailed studies to be carried out on different statistical, imaging and probabilistic scales of information, condensed in the so-called data representation systems or dataset. Consequently, there are countless applications in the health area at different levels, optimizing early diagnosis processes, minimizing the risks associated with a new global pandemic due to COVID-19 and / or any other virus.

II. COVID-19

A. Background

Coronaviruses (CoVs) belong to the Orthocoronavirinae subfamily of the Coronaviridae family in the order Nidovirales, and this subfamily includes α-coronavirus, βcoronavirus, γ-coronavirus, and delta-coronavirus (Banerjee et al., 2019). Among the causative agents of human respiratory tract infections are coronaviruses (CoVs), which are enveloped single-strand RNA viruses that belong to the large Coronavirinae subfamily that infect birds and mammals (Raoult et al., 2020). Regarding the new SARS coronavirus-2 (SARS-CoV-2) Guan et al. (2020) point out that it seems highly transmissible from a human-to-human pathogen, causing a wide spectrum of clinical manifestations in COVID-19 patients. Regarding the transmission of the virus, Yang and Wang (2020, 2710) affirm that, in the review of 22 types of coronavirus, both SARS-CoV, MERS-CoV and endemic human coronaviruses can persist on inanimate surfaces such as metal, glass or plastic for up to nine days, providing strong evidence of the environmental survival of the pathogen. Added to this evidence is the contamination of the water by feces of infected people, expanding another possible route of transmission of this disease. Regarding the inactivation of coronavirus by disinfectant agents in suspension tests, consult Kampf (2020) in which evidence is presented in this regard increased

B. COVID-19 Charichtaristics

This virus is an enveloped RNA virus that causes severe respiratory failure and belongs to the Korana virus family such as SARS-CoV and MERS-CoV, identified by determining human transmission on January 7, 2020. On December 31, 2019 on the same day, the WHO published a series of provisional guidelines for all countries on how to prepare for the possible arrival of this virus, regarding the way to control sick people, the analysis of samples, the treatment of patients and the control of infection in health centers. However, the trajectory of this outbreak was impossible to predict and, despite the implementation of classic public health strategies in many countries, the WHO Emergency Committee declared it on January 30 as a Public Health Emergency of International Importance (ESPII), that is, “an extraordinary event that constitutes a risk to the public health of other States due to the international spread of a disease, which may require a coordinated international response”. This ESPI statement implied, as early as the end of January, that the situation was: serious, unusual or unexpected, had implications for public health beyond the borders of the affected State, and could require immediate international action. Its main objective was to guarantee health security through the application of the Health Regulations International.3 3 Legally binding international agreement signed in 2005 by 196 countries, including all WHO Member States. Its objective is to help the international community to prevent and respond proportionally to the serious risks to public health arising from international spread.In spite of everything, during the months of February and March the epidemic spread rapidly, with dramatic increases in the number of number of infections and deaths from the disease, and with an important confirmed community transmission in many countries in Europe and other continents, which led the WHO to classify the disease caused by the new coronavirus on March 11 as a pandemic. From that moment, in Spain urgent measures of public health were established before the escalation of infections; For example, on March 12, in line with the recommendations of the European Center for Disease Prevention and Control (ECDC), social distancing measures were extended to the entire country. On March 14, the Government approved Royal Decree 463/2020 that declared a State of Alarm, with drastic measures to protect the health of citizens, contain the spread of COVID-19 and strengthen the National Health System; Among them, he highlighted the limitation of mobility and free movement of people, home confinement and the closure of various economic or educational activities [4].

III. MACHINE LEARNING

Machine learning is an AI discipline that uses algorithms to identify patterns, make predictions, learn from data, and make decisions. For In the case of COVID-19, machine learning is used for the diagnosis and identification of the population that is at greater risk of infection. It is also used for faster drug development, including the study of reuse of drugs that have been proven to treat other diseases. To do this, knowledge graphs are constructed and perform predictive analysis of interaction between drug and viral proteins (Zhou, Park, Choi & Han, 2018) and virus-host interactomes (Yang et al., 2019), predictive protein folding (Ivankov & Finkelstein, 2020), understanding of the Molecular and cellular dynamics of the virus, prediction and spread of a disease based on patterns, and even predicting an upcoming zoonotic pandemic. To conduct AI studies using machine learning (which includes deep learning in some cases), certain algorithms are required, such as decision trees, regression for statistical and predictive analysis, generative adversary networks, instance-based clustering, Bayesians, neural networks , etc. These algorithms use data science in which various mathematical calculations are performed, where the information density is broad, complex and varied. For example, find antiviral molecules (Ahuja, Reddy & Marques, 2020) that fight COVID-19 and identify millions of antibodies for the treatment of secondary infections (Ciliberto & Cardone, 2020). Another type of machine learning application revolves around the prediction of infection risks, based on specific characteristics of a person, such as age, geographical location, socioeconomic level, social and hygiene habits, pre-existing conditions and human interaction, among others. With these data, a predictive model can be established on the risk that an individual or group of people can bring to contract COVID-19 and factors associated with developing complications (Jiang et al., 2020) and even predict the results of a treatment.

A. *Deep Learning*

Deep learning is a subfield of machine learning, which seeks to classify data using correlational algorithms. It is based on certain neural network architectures, which allow you to hierarchize information (visual, auditory and written) through a segmentation of patterns categorized by levels. Under this criterion, learning takes place in stages, in a manner equivalent to what happens in a human. That is, you start with basic data and as more complex levels of them are scaled, you learn With regard to the COVID-19 pandemic, the global health system proved not to be able to carry out diagnostic tests in the short term, added to economic, logistical, technological infrastructure problems and lack of hospital staff. AI is helping to minimize these problems through the use of deep learning techniques, through image recognition for radiodiagnostic tests that, unlike tests standard clinics, gives results in a few minutes, and in them it is inferred whether or not a patient's lungs are sick with pneumonia specifically associated with COVID-19.

B. *Big data*

The volume in the field of Big Data demands large information processing and storage resources, which are represented in the "Variety" of data, which can be structured and unstructured. With regard to "Speed", it refers to the amount of data that is generated periodically and requires a Scalable technological infrastructure that allows its availability and access at any time. Regarding the "Veracity" and "Value", it is essential that the stored data be truthful, otherwise valuable computational resources would be wasted in unreliable or useless information, leading to incorrect results and decision making. On the "value", it is understood in the sense of extracting relevant information to define strategies and decision making. For the particular case of COVID-19, a large amount of data is generated that, when using AI to analyze them, allows differentiating families, treatments, risks, etc., which converges to reduce costs in the diagnosis and treatment of a patient, saving thousands of lives in the process. Therefore, assertively using the five adjectives mentioned must guarantee truthful and reliable information to be able to implement them in deep learning AI systems, machine learning or both

IV. PROPOSED METHOD

In all countries and regions, the number of total infected is unknown in real time, only officially confirmed cases are observed, and only after time can an estimate of total cases be obtained, never their daily evolution. So the only thing that can be done is to think that the confirmed is a proportion of the total that remains approximately constant in time. Therefore, it must be borne in mind that if the number of tests carried out increases ostensibly, as has actually happened in many parts, the appropriate corrections should be made to normalize the data as seen in the previous section. The confirmed contagion rate is defined as the quotient between new confirmed and accumulated cases up to the previous day, which uses the following equation:

$$P(t) = \frac{1}{1+e^{-at+b}}$$

Obviously, under this assumption of proportionality of the infections observed over the total, the confirmed rate coincides exactly with the contagion rate and can be modeled as a linear regression on the inverse transformation of the hyperbolic cosecant defined in the equation 20, taking into account that in each region can vary both the parameters of the regression and its variance as well as the instant of disruption. In this way, a smoothed forecast of the contagion rate is obtained with a point of disruption.

This first modeling approach will give us a very smooth curve in time, in fact it is continuous and in nicely differentiable around which the actually observed number of confirmed cases will move. From this difference between the observed value and the predicted one, it has been found that its normality can be accepted under a certain instantaneous Box-Cox transformation [Box and Cox (1964)], which can vary in each region. It has also been empirically verified that although it is stationary3 it is not independent of the past, as can be seen in its temporal autocorrelation function

It is therefore a nested model that assembles the output of a deterministic cut model with a stochastic model with a linear dynamic structure and different degrees of autoregressive p and of moving average q according to each region, which have been automatically identified using an algorithm based on the Bayesian information criterion and a battery of diagnostic tests on autocorrelations, normality, significance and stationarity:

$$P(t) \;=\; \frac{M}{1+e^{-at+b}}$$

Given the great uncertainty that affects all these processes, it is inevitable that the sum of the confirmed forecasts in each region independently does not exactly coincide with the aggregate total (with g = 0), although as observed in the figure 10 there is not a big discrepancy either. In any case, it is convenient to combine the forecasts statistically to increase the robustness of the system and to ensure that the results are consistent with each other. an extra parameter α is added as follows:

$$P(t) \;=\; \frac{M}{\left(1+e^{-at+b}\right)^{\alpha}}.$$

The procedure followed is simply to weight each forecast inversely proportional to the variance of the errors made by each model where y = x 1/3, y = x 1/2, y = x, y = x 2, y = x 3. For this function the inflection point is when:

$$P''(t) \;=\; a^2ce^{at+b}(e^{at+b}+1)^{-c-2}(ce^{at+b}-1)$$

so, we have that the inflection point is obtained for:

$$t=-\frac{\ln(\alpha)+b}{a}$$

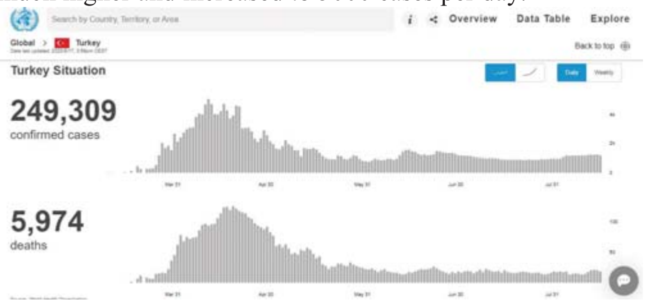
All the deceased have been confirmed at an earlier time or on the same day if the case is confirmed in the death report itself. Thus, there exists a series of numbers which we will call mortality transfer coefficients, such that. The case:

$$Y(t)=\frac{K}{(1+Qe^{-\alpha\nu(t-t_0)})^{1/\nu}}$$

which is a solution of the differential equation:

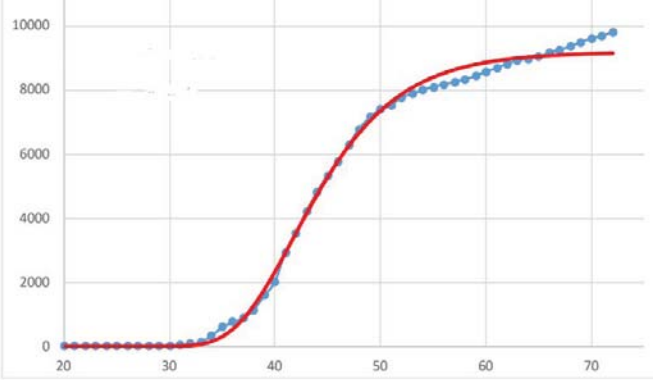
$$Y'(t)=\alpha\left(1-\left(\frac{Y}{K}\right)^{\nu}\right)Y$$

For the recovered there is another series of numbers 0 ≤ γ r g, k ≤ 1 analogous to the previous one, which we will call recovery transfer coefficients. Taking into account that in order to be discharged, a minimum quarantine period must be passed, let's say Q days, in this case we have to Let us now look at the case of Turkey where the data we got is from Belkent University (https://library.bilkent.edu.tr/covid-19-trial-databases/), in which the bend is practically finished. As we see when utilizing the days from 16 to 45 (20 days), note that the initial 16 days there were practically no adjustments in the information so information was not utilized, with these information with LG you get a most extreme blunder of 9.91% is acquired (normal of 3 runs, essentially gives a similar worth) When utilizing the days from 15 to 55 (30 days) the chart of figure underneath is gotten, when estimating the relative mistakes a greatest estimation of 4.49%is obtained (average of 3 runs), in the prediction of the data until the 31 / 03/2020 where cases in Istanbul where much higher and increased to 5000 cases per day:



Covid-19 Cases in turkey from WHO website

For Iraq we are going to use the data from day 28 from the first case (02/16/202) when an Iranian student reached Najaf, the data collected are obtained from (https://coronavirus-covid-19-iraq-atlasgis.hub.arcgis.com/), in this case remember that these data have a linear trend in the last days, see figure below:



Data from IRAQ and (real and predicted).

V. Conclusion

With the use of disruptive technologies such as artificial intelligence and Big Data, it is expected to be better prepared for the next pandemic, even to prevent it. Technologies 4.0 such as the internet of things, smart computing and cloud computing will contribute their own in terms of permanent monitoring of cities in search of biological and chemical anomalies that imply some risk to society or the environment. This type of development can be personalized, since applying predictive learning algorithms minimizes risks when formulating treatments that can establish whether a patient tolerates them or not. What can be rescued from the COVID-19 pandemic is that it has promoted unprecedented technological developments in terms of artificial intelligence in its different areas of knowledge, just like big data science. Under this scenario, the health sector will have to quickly incorporate these resources into its analysis and diagnosis system, not only of infectious diseases but of any other, so it is expected to improve the service provided to a patient or community and prepare society in the event of a pandemic in the future. In addition, these types of developments help health centers to reduce operating costs of various kinds, where diagnosis time plays a fundamental role in stopping a potential pandemic outbreak. AI together with Big Data have proven to be fundamental tools to help the health sector to detect and control this virus with a certain margin of success, allowing the processing of large amounts of structured and unstructured data with a high degree of complexity, which when combined With AI algorithms, they allow predictions based on historical patterns and feedback loops, among others. What is important about this synergy is that it helps medical care more effectively, even after the crisis is over. Also, with the learning that is constantly being developed, there are already developments of predictive algorithms that allow identifying populations that are or will be more likely to be infected by COVID-19, even determining in probabilistic terms who may suffer serious complications based on parameters such as age, gender, medical history, body mass, among others. With this type of development, these algorithms can be extended to be applied to other types of diseases, thus contributing to improve the health service. AI and radiodiagnosis are playing an important role in the detection of COVID19 with a percentage higher than 90%, which can be increased when the system with the largest amount of data is trained, therefore Big Data in conjunction with other analytical disciplines they are a key factor in successfully completing a study. It should be noted that AI does not replace the health professional, on the contrary, it is a complement to their medical work, helping them to improve the accuracy of the diagnosis in a shorter time and make decisions much faster, thereby lightening their workload

VI. REFERENECES

[1] Alsulaiman, M., Alotaibi, Y., Ghulam, M., Bencherif, M. A., & Mahmood, A. (2010). Arabic speaker recognition: Babylon Levantine subset case study. Journal of Computer Science, 6, 381– 385. doi:10.3844/jcssp.2010.381.385

[2] Alsulaiman, M., Ghulam, M., Bencherif, M. A., Mahmood, A., & Ali, Z. (2013). KSU rich Arabic speech database. Information Journal, 16, 4231– 4254. http://catalog.ldc.upenn.edu/LDC2014S02

[3] Altıncay, H., & Demirekler, M. (2002). Why does output normalization create problems in multiple classifier systems? In Proceedings of CPR2002, 16th International Conference on Pattern Recognition, Quebec, Canada.

[4] Anusuya, M. A., & Katti, S. K. (2011). Front end analysis of speech recognition: A review. International Journal of Speech and Technology, 14, 99 – 145. doi:10.1007/s10772-010-9088-7

[5] Fukuda, T., & Nitta, T. (2003). A study on Japanese distinctive phonetic feature set for robust speech recognition. Autumn Meeting of the Acoustical Society of Japan, I, 9 – 10 (in Japanese).

[6] Fukuda, T., & Nitta, T. (2004). Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition. IEICE Transactions on Information and Systems, E87-D, 1110– 1118.

[7] Gonzalez-Rodriguez, J. (2014). Evaluating automatic speaker recognition systems: An overview of the NIST speaker recognition evaluations (1996 – 2014). Loquens, 1, e007. http://dx.doi.org/10.3989/ loquens.2014.007

[8] Hassan, F., Kotwal, M. R. A., Rahman, M. M., Nasiruddin, M., Latif, M. A., & Huda, M. N. (2011). Local feature or mel frequency cepstral coefficients – Which one is better for MLN-based Bangla speech recognition? Springer-Verlag B, 2011, 154– 161, Berlin.

[9] Jayanna, H. S., & Mahadeva Prasanna, S. R. (2009). Analysis, feature extraction, modelling and testing techniques for speaker recognition. IETE Technical Review, 26, 181– 190.

[10] Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Communication, 60, 56 – 77. doi:10.1016/j.specom.2014.03.001

[11] Lawson, A., et al. (2011), Prague, Czech Republic Survey and evaluation of acoustic features for speaker recognition. ICASSP.

[12] Li, K. P., et al. (1966). Experimental study in SV using adaptive system. JASA, 40, 1441– 1449.

[13] Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martı´nez-Gonza´lez, D., Gonzalez-Rodriguez, J., & Moreno, P. J. (2014). Automatic language identification using deep neural networks. IEEE International Conference on acoustics, speech and signal processing (ICASSP ’14), Florence, Italy.

[14] Mahmood, A., Alsulaiman, M., & Muhammad, G. (2012). Multidirectional local features for speaker recognition. ISMS 2012.

[15] Kota Kinabalu, Malaysia. Nitta, T. (1998). A novel feature-extraction for speech recognition based on multiple acoustic-feature planes. Proceedings of the IEEE ICASSP’98, I, 29 – 32.

[16] Nitta, T. (1999). Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA. Proceedings of the IEEE ICASSP’99, I, 421–424.

[17] Qiu, A., Schreiner, C., & Escabi, M (2003). Gabor analysis of auditory midbrain receptive fields: Spectrotemporal and binaural composition. Journal of the Neurophysiology, 90, 456– 476. doi:10.1152/jn. 00851.2002

[18] Reynolds, D. (1995). Large population speaker identification using clean and telephone speech. IEEE Signal Processing Letters, 2, 46 – 48. doi:10.1109/97.372913

[19] Reynolds, D. A., Quatieri, T. F., & Dunn, R (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10, 19 – 41.