



T.C.

ALTINBAS UNIVERSITY

Institute of Graduate Studies

Information Technologies

**PREDICTING COVID-19 CASES TRAJECTORY
USING MACHINE LEARNING**

Zainab Abbas Abdulhussein ALWAEI

Master of Science

Supervisor

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Istanbul, 2021

PREDICTING COVID-19 CASES TRAJECTORY USING MACHINE LEARNING

by

Zainab Abbas Abdulhussein ALWAEI

Information Technologies

Submitted to the Institute of Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAŞ UNIVERSITY

2021

The thesis titled “PREDICTING COVID-19 CASES TRAJECTORY USING MACHINE LEARNING” prepared and presented by “Zainab Abbas Abdulhussein ALWAEI” was accepted as a Master of Science Thesis in Information Technologies.

Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Supervisor

Thesis Defense Jury Members:

Asst. Prof. Dr. Abdullahi Abdu
IBRAHIM

School of Engineering and
Architecture,

Altinbas University

Asst. Prof. Dr. Sefer KURNAZ

School of Engineering and
Architecture,

Altinbas University

Prof. Dr. Adem KARAHOCA

Computer Engineering,

MEF University

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Approval Date of Institute of Graduate Studies:

____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Zainab Abbas Abdulhussein ALWAEI

DEDICATION

I would like to thank Allah Almighty for the power of mind, health, strength, guidance, knowledge and skills to complete this study.

This thesis is wholeheartedly dedicated to my parents. There are no words to describe what you mean to me; there is nothing that I can repay for what you have done to me. I will continue to do my best to achieve your expectations.

Lastly, I dedicated this to my wife, brother and sisters, relatives, and friends who have been encouraging me during this study.



ACKNOWLEDGEMENTS

I would like to thank very much my supervisor, Asst. Prof. Dr. Abdullahi Abdu IBRAHIM for his assistance throughout the research period and for answering many questions and he was the best help for me throughout the study period. Without his time and help, this research would not have been completed and I am very grateful to him.



ABSTRACT

PREDICTING COVID-19 CASES TRAJECTORY USING MACHINE LEARNING

Zainab Abbas Abdulhusssein ALWAEI

M.Sc., Information Technologies, Altinbas University,

Supervisor: Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Date: 09/2021

Pages: 46

Smart grids are electric grids that are composed of multiple power sources and devices connected to each other to provide better reliability in power generation and power management, modern developments of the smart grid aim at either improving the control of power sources and loads connected to the smart grid by developing a specialized software/hardware, or by improving the communication within the parts of the smart grid and the central control. In this paper we aim at improving both sides of the smart grid system (communication and control), we propose a fuzzy logic based controller for renewable energy and fossil fuel sources in a grid and an internet of things based monitoring system which oversees the state of the smart grid, faults that occur in the grid, and how the fuzzy controller overcomes those faults, all in which provide an extra layer of support to the smart grid.

Keywords: Photovoltaic, Solar energy, Fuzzy logic, Smart grid, Internet of things IOT, MATLAB Simulink.

TABLE OF CONTENTS

	<u>Pages</u>
ABSTRACT.....	vii
LIST OF FIGURES	x
1. INTRODUCTION.....	1
1.1 BACKGROUND	1
1.2 MOTIVATION	1
1.3 PROBLEM STATEMENT.....	2
1.4 CONTRIBUTION.....	2
1.5 THESIS ORGANIZATION	3
2. LITRITAURE REVIEW	4
3. MATERIALS AND METHODS	5
3.1 COVID-19	5
3.1.1 COVID-19 Charcteristics	7
3.1.2 COVID-19 Diagnostics	8
3.2 ARTIFICIAL INTELLIGENCE.....	9
3.3 MACHINE LEARNING	11
3.3.1 Machine learning Algorithms	12
3.3.2 Deep learning	13
3.4 BIG DATA	16
3.5 ETHICAL ASPECTS OF USING MACHINE LEARNING.....	18
4. PROPOSED METHOD	19

4.1	MATHEMATICAL MODELLING.....	19
4.2	LOGISTIC REGRESSION	20
4.3	ADJUSTING THE CURVE	21
4.3.1	China Data	21
4.3.2	South Korea Data	22
5.	SIMULATION AND RESULTS	24
5.1	TURKEY DATA	24
5.2	IRAQ DATA.....	25
6.	CONCLUSION	31
	REFERENCES.....	32

LIST OF FIGURES

	<u>Pages</u>
Figure 1.1: Countries effected by Covid-19[1].....	1
Figure 1.2: General Covid-19 prediction model[2]	2
Figure 3.1: Distibution of COVID-19 cases through turkey[3].....	6
Figure 3.2: Microscopic images of Covid-19[4]	7
Figure 3.3: Lung damage caused by COVID-19[5].....	8
Figure 3.4: Types of COVID-19 tests[6].....	9
Figure 3.5: Types of AI[7].....	10
Figure 3.6: Machine learning subgroups[8]	11
Figure 3.7: Hierarchy of AI[9].....	14
Figure 3.8: Evolution of big data[10].....	17
Figure 4.1: China data	22
Figure 4.2: S-Korea data	23
Figure 5.1: Covid-19 Cases in turkey WHO website.....	24
Figure 5.2: Data from IRAQ and (real and predicted).....	25
Figure 5.3: Population growth curves can be adjusted, using the LG 1	25
Figure 5.4: Population growth curves can be adjusted, using the LG 2	26
Figure 5.5: Population growth curves can be adjusted, using the LG 3	27
Figure 5.6: Population growth curves can be adjusted, using the LG 4	28
Figure 5.7: China population related to Covid-19 cases	29
Figure 5.8: Turkey Covid-19 cases and population density map	30

1. INTRODUCTION

1.1 BACKGROUND

The pandemic caused by COVID-19 in 2020 (Sohrabi, et al., 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:

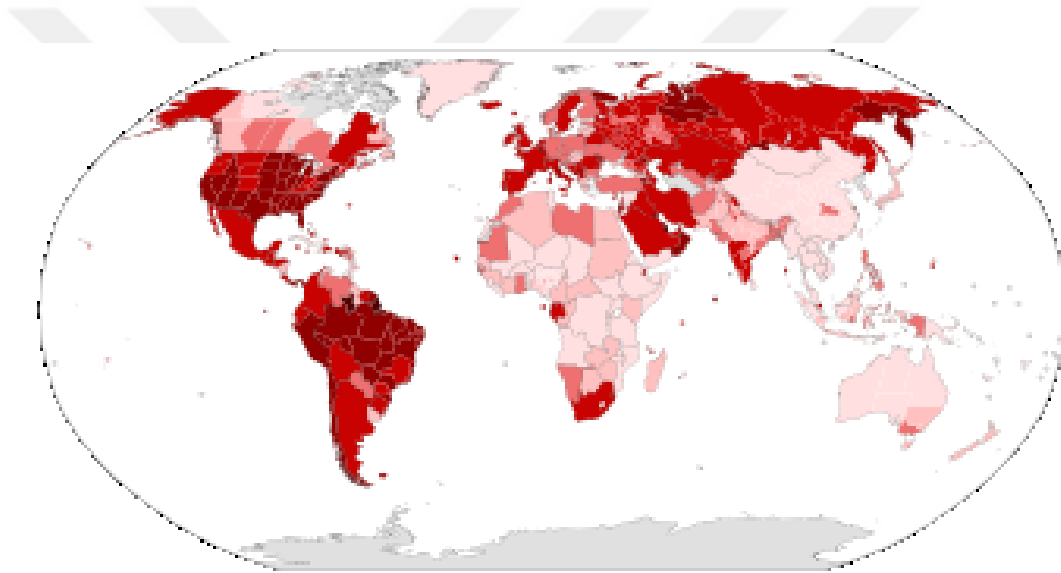


Figure 1.1: Countries effected by Covid-19 [1]

1.2 MOTIVATION

Although the techniques are refined, their sensitivity to samples is relative. The problem arises in the procedure for collecting the tests that is carried out manually, to which is added the enormous amount of information that needs to be processed, such as data related to the rapid transmission, molecular and cellular dynamics of the virus, traceability about population and ethnic susceptibility associated with the pandemic, including monitoring the level of risk of employees in a company, among other aspects.

1.3 PROBLEM STATEMENT

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:

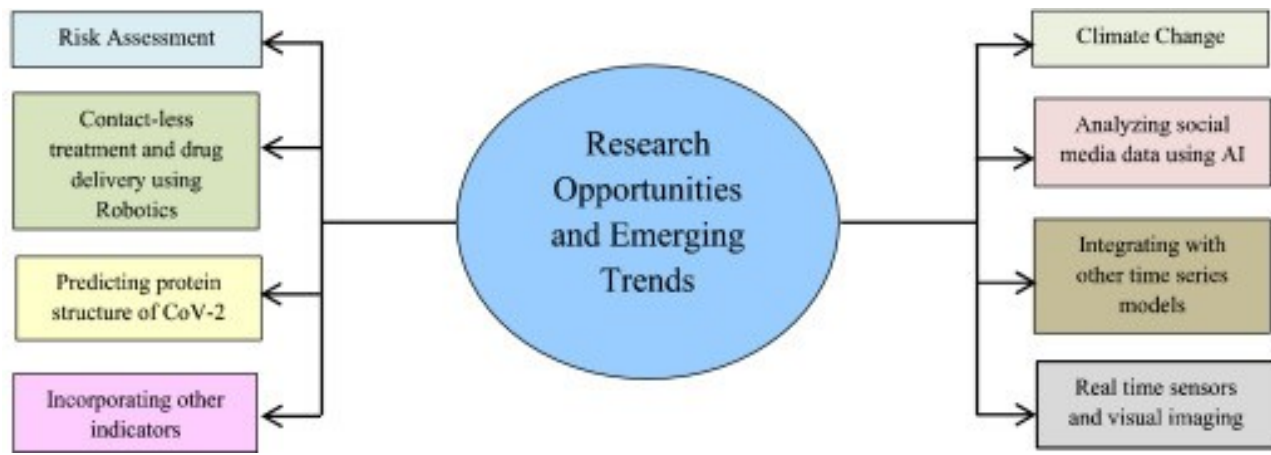


Figure 1.2: General Covid-19 prediction model [2]

1.4 CONTRIBUTION

In this work 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.

1.5 THESIS ORGANIZATION

The thesis is structured as follows: Section 2 is where we review some of the previous work and implementation on smart grids. Section 3 is where we give a background about all the components. Section 4 is where we explain our model in details and implementation of that model in details, Section 5 is where we simulate our model and record the results obtained. Section 6 is where we conclude our work and put a future scope into perspective.



2. LITERATURE REVIEW

John Daugman, describes in [9] and [10], a methodology for the process of translating Cartesian coordinates to the polar coordinate system of non-concentric circles, called the rubber sheet.



3. MATERIALS AND METHODS

In this chapter we review some of the essential components that we used to implement our model, we give a brief introduction to the component, and then we show how we simulate that component inside MATLAB environment.

3.1 COVID-19

Coronaviruses (CoVs) belong to the Orthocoronavirinae subfamily of the Coronaviridae family in the order Nidovirales, and this subfamily includes α -coronavirus, β -coronavirus, 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

2019 YIL SONU İTİBARIYLA İLLERE GÖRE YAKALANAN DÜZENSİZ GÖÇMENLER



Figure 3.1: Distribution of COVID-19 cases through turkey [3]

Based on these facts, COVID-19 has brought to the fore the fragility of humanity in the face of a virus that has proven to be highly resistant and dangerous, with an unprecedented spread factor on a global scale in modern times, exposing the lassitude of health sector in attending a growing wave of infected that amounts to millions and fatalities that add up to hundreds of thousands. Although faster diagnostic techniques (Pang, et al., 2020) and experimental antiviral drugs and therapies are being developed, challenges remain to overcome such as speeding up manufacturing processes and overcoming problems related to the supply and supply chain (Ledford, 2020). Meanwhile, the early diagnosis and specific treatment of this disease present a series of difficulties that have prompted the academic and research communities to formulate proposals framed from the perspective of disruptive technologies such as artificial intelligence and massive data processing, which allow finding alternative solutions on prevention, control, replication and spread of the virus, which many scientists claim can become endemic (Abdessater et al. 2020; Kampf et al. 2020).

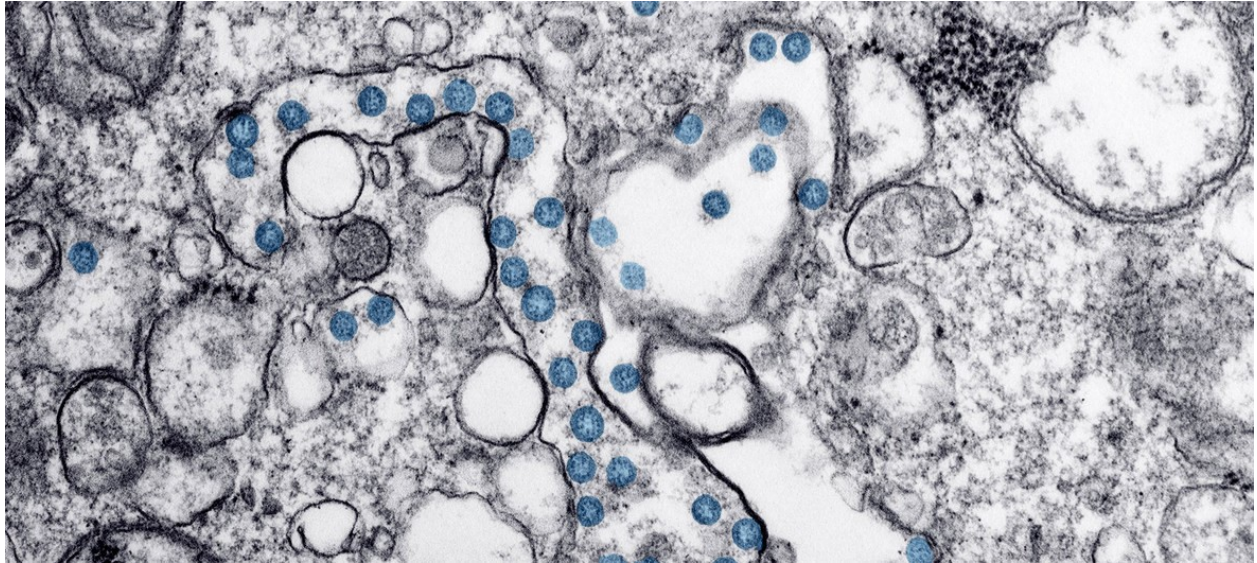


Figure 3.2: Microscopic images of Covid-19 [4]

Given the above scenario, technology has come to demonstrate its worth in helping to find and establish certain factors related to COVID-19, as well as in establishing its molecular and cellular behavior, as well as in the way of spreading and infecting an individual, among others. factors. That is why various research groups, institutions, companies and governments have proposed initiatives to seek short-term solutions to contain the pandemic generated by the virus, as well as to find a cure against it.

3.1.1 COVID-19 Characteristics

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, it was determined that a case reported from Wuhan city with a population of 11 million in Hubei province in China was infected with a new Coronavirus that has never been seen before. This virus is also thought to be transmitted from animals to humans, a zoonotic infection such as SARS-CoV and MERS-CoV. Covid-19 was first seen in people in the seafood market in Wuhan, China, in December 2019. Although scientists thought that Covid-19 was transmitted from bats, it was suggested that a type of intermediate animal was a carrier in the transmission of the virus to humans, considering that bats were not sold in the market. Fast today the disease has spread from person to person, spreading to Asia, Europe and the Americas other than China, with the number of patients reaching millions and the number

of deaths reaching tens of thousands. The outbreak still continues at the pandemic level (as of April 2020). According to the researchers, this virus can survive on different materials such as aluminum, wood, paper, plastic and glass for a maximum of 4-5 days. According to a study published in The Lancet magazine in 2019, the most common symptoms of Covid-19 were determined to be fever, cough, and shortness of breath. Sore throat was observed in 5% and diarrhea, nausea and vomiting in 1-2%. Pneumonia, severe acute respiratory failure, multiple organ failure and death have been observed in patients with advanced course. It has been determined that the virus is transmitted from person to person, usually through droplets formed as a result of cough. Contamination has also been observed by touching one's own face after touching contaminated surfaces. Since there are virus particles in the stools of the patients, there may be fecal-oral transmission. The incubation period is 2-4 and the disease duration is determined as 5 days.



Figure 3.3: Lung damage caused by COVID-19 [5]

3.1.2 COVID-19 Diagnostics

Various immunological tests and Polymerase Chain Reaction (PCR: Polymerase Chain Reaction) studies performed by targeting specific genes are of great importance in the diagnosis of the virus. The most ideal diagnosis method of Covid-19; Quantitative Real by nasopharyngeal swab Time Polymerase Chain Reaction (qRT-PCR) method. For this method, sensitive and specific (sensitive and specific) results are obtained by using oligonucleotide probes labeled with fluorescent dyes, one of the unique detection systems. With this method, symptoms, risk factors and signs of

pneumonia should be evaluated together with thoracic CT scans in the diagnosis of the disease. Although the sensitivity of PCR tests is high, antibody tests (immunoassays) that give rapid results have been produced due to the need for experienced personnel and taking the test result in 4 - 6 hours (reduced to 2-4 hours in the final kits). Tests have been produced that can detect IgM and IgG antibodies produced against SARS-CoV-2 in 15 minutes. However, there is a 13% chance of giving false negatives in catching positive cases. A rapid diagnostic kit has been developed for the detection of coronavirus in our country. This kit developed has been used routinely in the National Virology Laboratory. Different diagnostic kits have also been developed in the process.

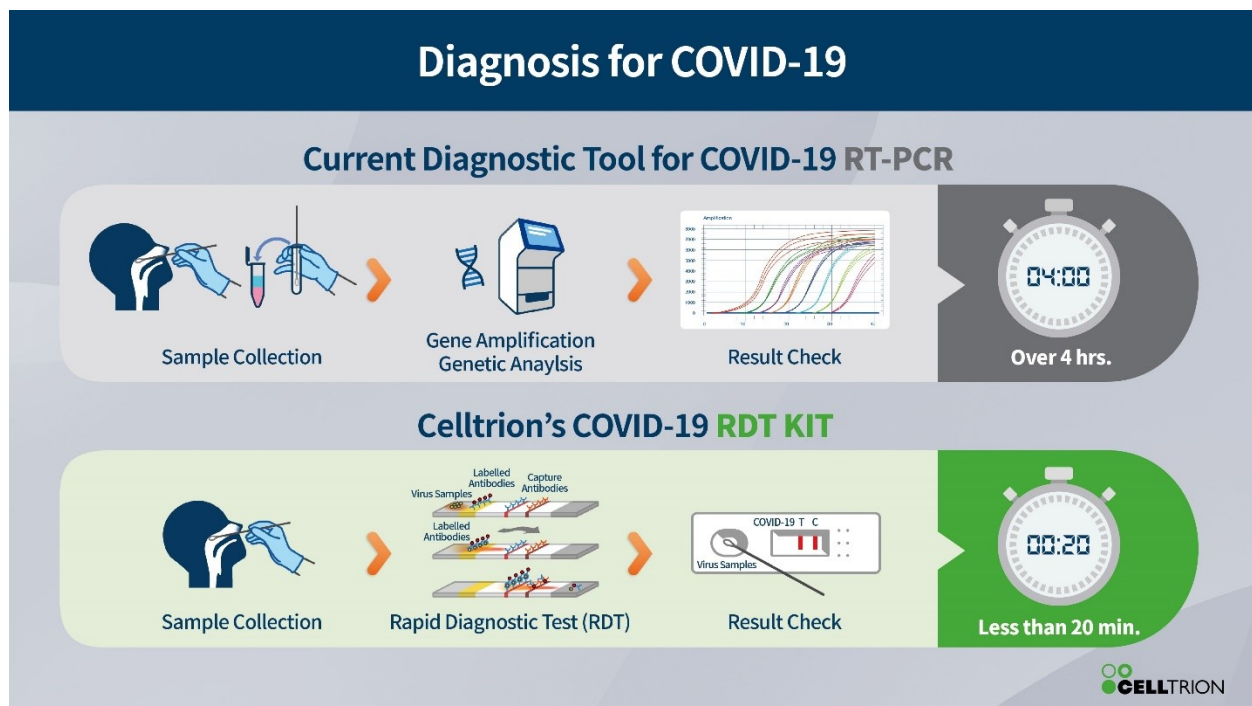


Figure 3.4: Types of COVID-19 Tests [6]

3.2 Artificial intelligence

Artificial intelligence or AI can be understood as a discipline belonging to computer science, which raises computational learning models based on human biological neural networks. In this sense, various AI models have been proposed, which thanks to advances in computer technology have allowed the development of "intelligent" systems that facilitate the processing of more data in less time, speeding up decision-making. One aspect to mention about AI is that it covers various fields such as speech recognition, natural language processing, computer vision, robotics advanced, knowledge capture, planning and optimization, among others, in which it is sought that a system has the capacity to feel, reason, participate and learn. As a complement to these fields, there are

specialties such as machine learning (Machine Learning) that involves different types of neural networks encoded under reinforcement learning, supervised learning, unsupervised learning, and semi-supervised learning, which converge in systems that work with advanced algorithmic techniques to process various types of signals, optimizing models, processing massive information through Big Data and Clusters; all this through the use of algorithmic regression, classification and clustering, among others. Another specialty of AI is deep learning (Deep Learning) to evaluate image, video and audio data using convolutional neural networks with their various variants (Tang et al., 2020) and short and long-term memory neural networks. Too allows the simulation of dynamic systems through the use of reinforcement learning. Based on the behavioral dynamics of COVID-19, prompt solutions are required for the monitoring, detection and diagnosis of the diseases generated by its cause (Law, Leung & Xu, 2020), the AI raises various hardware and software options aimed at this end (Mei et al., 2020). Under this scenario, the development of open source software has been gaining strength, where collective intelligence is the main gear to obtain a high-performance program, multipurpose in most cases.

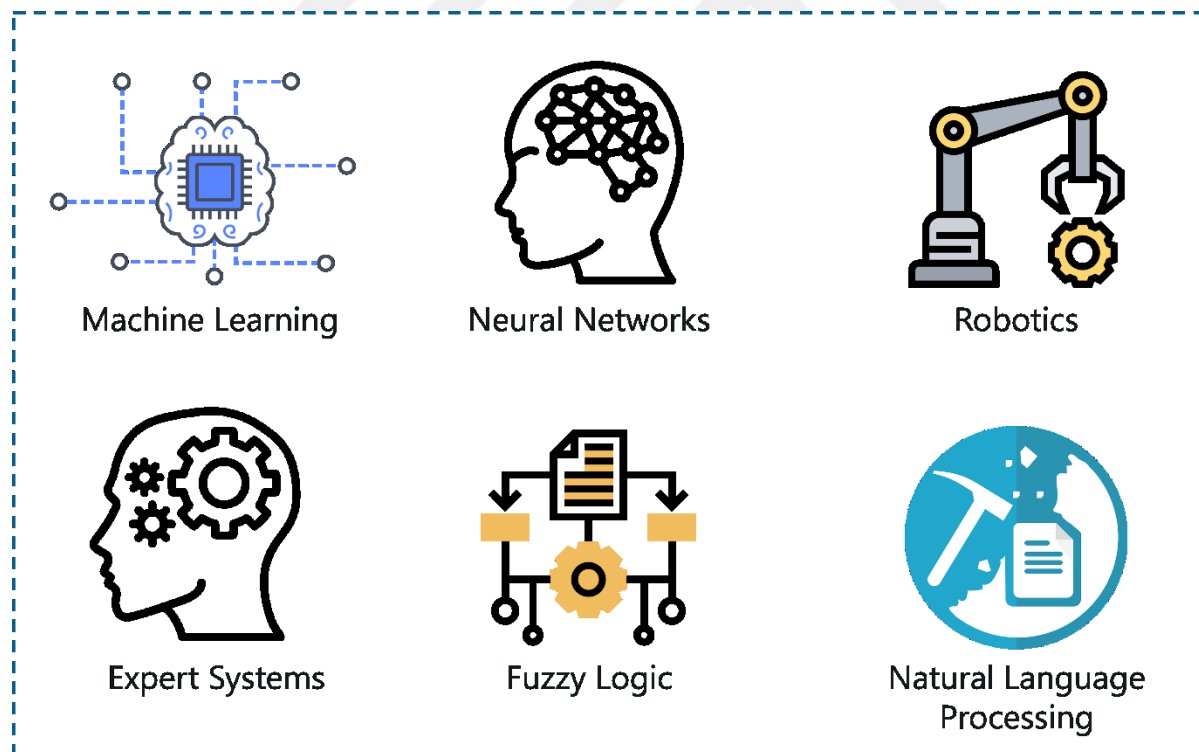


Figure 3.5: Types of AI [7]

3.3 Machine Learning

One particularity of working with collective intelligence lies in the fact that it is aimed at researchers interested in helping to improve the system, perfecting it until reaching a functional, reliable and one hundred percent operational product. An example is the CovidNet system, whose objective is the detection of viral infection by COVID-19 by examining chest x-rays or tomographies (Wang & Wong, 2020), where various studies allow identifying radiographic abnormalities in images with a high degree of reliability, such as viral pneumonia and other respiratory problems that may or may not be related to the virus. In the same way, AI is being used to discover compounds that derive into drugs, not only to deal with the coronavirus, but to optimize the processes and procedures carried out in the pharmaceutical and laboratory industries, where the use of Dataset is fundamental, since they contain huge collections of updated information, generally expressed through Data Tables and Data Relations, which allow to filter massive information more quickly. With this in mind, research and development costs are significantly reduced, being reflected in the final value of the drug for the patient and the health sector in general. The tools to be used for this type of case are generative models, which are based on machine learning and deep learning, which allow working with a large amount of data related to the molecular dynamics and physicochemical properties of a drug, facilitating the identification of thousands of new candidate molecules to be synthesized and tested.

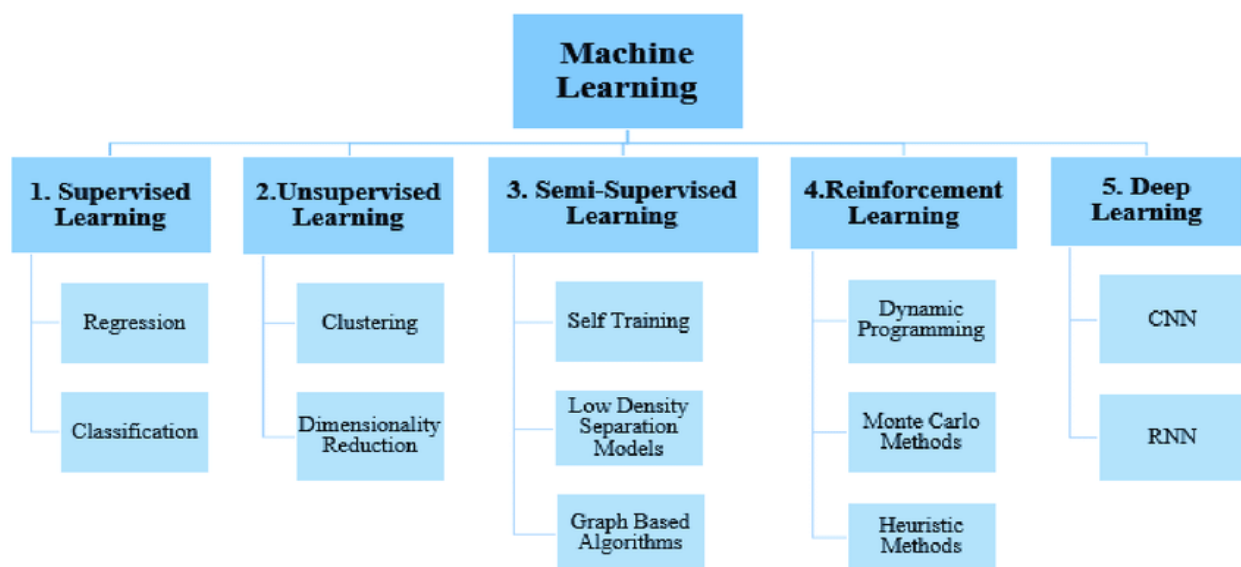


Figure 3.6: Machine learning subgroups [8]

3.3.1 Machine learning algorithms

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. With these types of projections, you could literally predict whether a patient lives or dies. Machine learning opens up a myriad of research possibilities in various clinical fields, where COVID-19 has been the driving force behind it. This involves everything from facial scanners to identify symptoms such as fever, wearables to measure and detect cardiac or respiratory abnormalities, to chatbots that evaluate a patient when he mentions his symptoms and, based on the answers given, the system tells him if you must stay home, call the doctor, or go to the hospital. The advantage of using machine learning over other standard techniques that take years is that the identification process can be completed in a matter of weeks, with a considerable cost reduction, coupled with a very high probability of success. For example, Smith and Smith

(2020, 2) state that the future design of SARS-CoV-2 antiviral drugs is already in the hands of a European team that uses the IBM supercomputer equipped with the AI SUMMIT system to be used in COVID-19 treatments.

3.3.2 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 (Öner, 2020). This specificity lies in the fact that through extraction processes of morphological characteristics on the object of study (Márquez, 2019), they are validated with respect to patterns of other diseases, establishing similarity factors, either in terms of the geometry of the tissue, injuries in the same and variations of intensity, among others. Deep learning is characterized by using graphic information to carry out its analysis. In the case of COVID-19, different types of algorithms have been developed aimed at the early detection of lung problems, either through X-ray images (Sánchez, Torres & Martínez, 2020), tomography or ultrasound (Fraile, 2020). Ongoing investigations They base their development on data science, because it provides analysis tools that allow proposing plausible solutions to deal with future pandemics as well as current treatments for COVID-19. For example, communities around the world have emerged that have created platforms that work under the collective intelligence model, where developers, data scientists, researchers, doctors, among other volunteers, formulate projects related to COVID-19. A platform that centralizes this information is Helpwithcovid, which has libraries such as the Covid Healthcare Coalition, which allows hundreds of thousands of researchers to access key information to work with deep learning. It is important to mention that the Republic of Korea provides access to anonymous information on patients with COVID-19, including the five-year medical history of

each patient, so that when conducting a deep learning traceability study there is a high amount of information useful for these purposes. To protect the privacy of the patient, those who execute the code are a specialized group, and then once processed they return the results to the researchers. This study is of special interest because it allows epidemiological monitoring, characterizing the population that falls ill, but does not enter a hospital, or those that never show symptoms. Under these circumstances, it is very difficult to track how the disease spreads and how many people it actually kills. Thus, with research using AI, it is possible to obtain predictive information that helps health authorities to take action on the matter. For example, using a model of branching to estimate how many people have been infected, viral DNA extracted from each known patient is analyzed, then the model uses the mutation rate to interpolate to how many others people passed the virus along the way (Li & Ayscue, 2020). Another type of AI research using deep learning is through the use of natural language, which allows creating new drugs from the search for existing compounds (Freedman, 2019). To do this, we analyze how the virus delivers its genetic material to an infected cell, in order to create a predictive model about the protein structure that allows posteriori to search for those that can inhibit this process (Réda, Kaufmann & Delahaye, 2020).

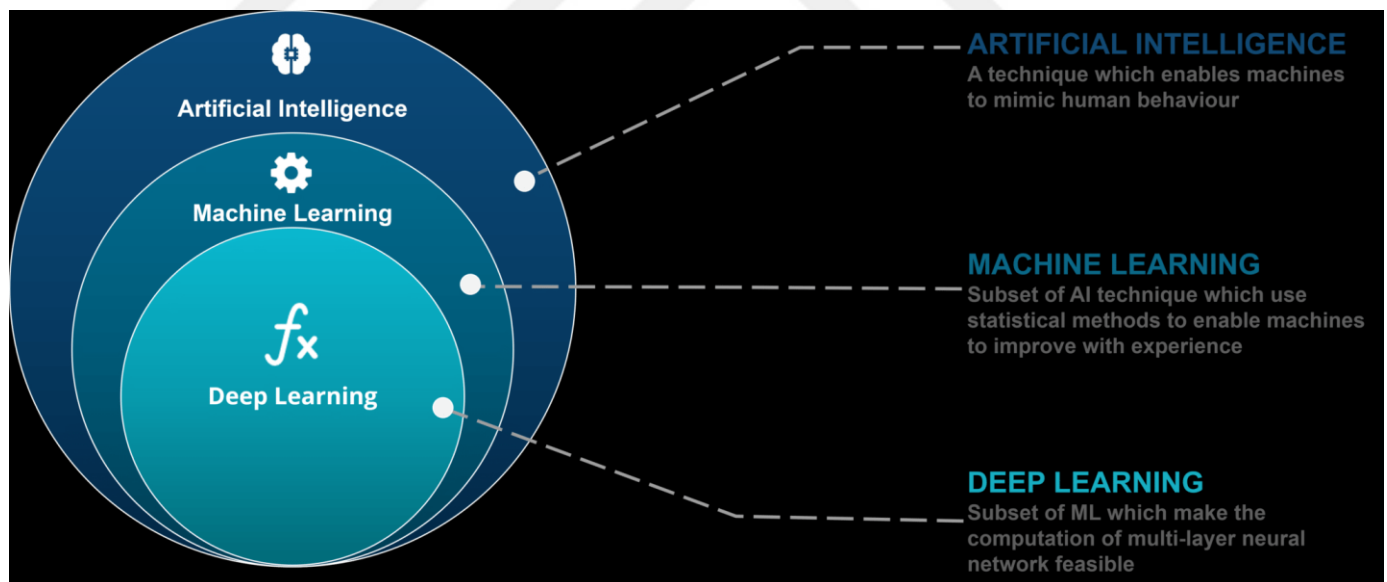


Figure 3.7: Hierarchy of AI [9]

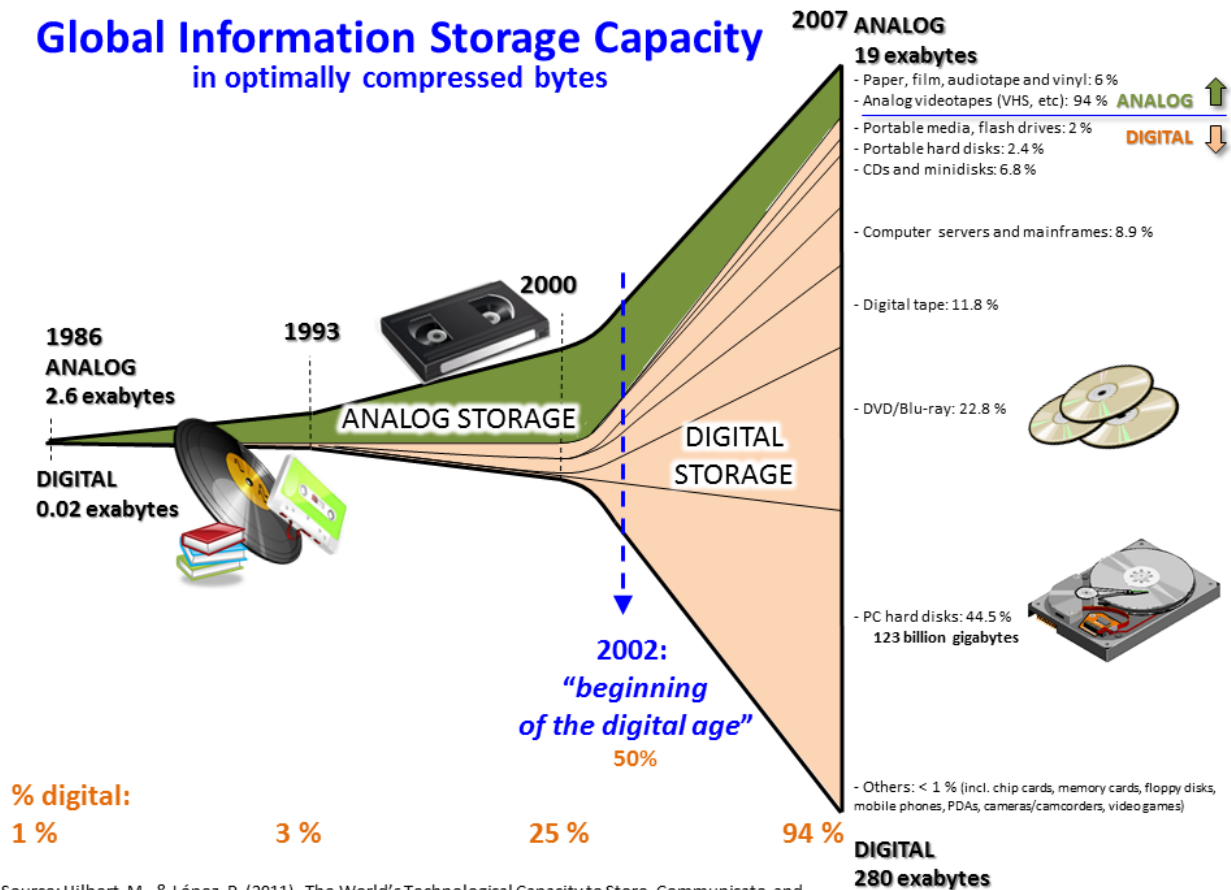
Other types of deep learning development, based on face recognition systems commonly used in physical security, have been modified to detect whether the community is complying with regulatory social distance. This software uses standard video cameras or those arranged in a city for video surveillance, allowing to monitor the pedestrian flow in critical areas, making a

recognition on the distance minimum, indicating an alert to the authorities if someone does not comply with the norm. In some cases, this system is combined with thermal measurement cameras in order to establish if any person has a fever and issue an alert to be isolated immediately by health personnel. In the clinical field, deep learning using images demonstrates robust results and allows the identification of various types of diseases such as cancer, glaucoma and pneumonia; the latter associated with COVID-19 under certain cell recognition patterns, with a very low level of false positives. For example, there are applications present in the market that use X-ray images for the early detection of Coronavirus, in which lung abnormalities associated with viral infection are identified (Ozturk et al., 2020) based on the rate of virulence and transfer detected, which with other techniques would take more time to specify. In general, creating an automatic system combining deep learning models facilitates the detection of various pulmonary anomalies and, additionally, the risk of disease can be calculated using probabilistic models from an X-ray. One particularity of the technologies associated with AI is that they require increasingly powerful algorithms that facilitate processing with a greater amount of information; which demands computational power with centralized services in the cloud. Based on this requirement, there are new algorithms that optimize certain models related to deep learning found in chips designed for this purpose. For example, the Intel chips (Nervana (NNP) and the NNP-I), the Nvidia family of AI chips, Google and Amazon chips, which promise to run AI with less power without sacrificing efficiency, even power. Perform inferences. These developments are known as tiny AI; whose particularity lies in the fact that various types of clinical image analysis can be performed directly from a mobile device. This allows for in situ studies on some type of histological dynamics related to COVID-19 or any other infectious agent through the use of imaging patterns. In the case of robust data analysis, a large technological infrastructure is required, such as supercomputers specifically dedicated to working with AI, where companies such as Microsoft and IBM are betting with the aim of finding treatments that allow to combat COVID-19 in a relatively long time short. The peculiarity of working with this type of technology is that it can be executed from different parts of the world, facilitating access to open source to groups of researchers working in various areas of AI, Big Data and data science.

3.4 BIG DATA

Big Data is understood as the set of computational procedures applied to analyze large amounts of data in order to extract information that presents certain patterns, relationships and associations relevant to an organization. Because more information is progressively created in different fields of knowledge, the trend for the coming years is the massive and recurrent integration of Big Data with the Internet of Things in an urban and industrial environment, where not only variables such as the air quality, relative temperature and humidity, but also integration with biometric systems, monitoring and surveillance cameras that seek to help the health sector by studying the health status of people in situ (preventive health) using sensors and thermal cameras, as well as the early identification of possible infectious foci through the use of biosensors and nanosensors strategically placed in the city. It should be noted that Big Data is characterized by three proper adjectives for the development of a research project, which are "Volume", "Variety" and "Speed", known as the 3 V of Big Data; other authors (Ishwarappa and Anuradha, 2015) add "Veracity" and "Value". 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.

Global Information Storage Capacity in optimally compressed bytes



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

Figure 3.8: Evolution of Big data[10]

Based on the above, the study of COVID-19 using Big Data can make use of advanced retrospective and descriptive analytics (Mohamed et al., 2019) expressed in Business intelligence (Chahal, Jyoti and Wirtz, 2019); since this allows to focus the study through indicators and trends over time, which includes future predictions. Under these criteria, data science becomes part using statistical and mathematical techniques characterized in the study variables that allow expanding the techniques and models represented as clusters (Mohebi et al., 2016) through patterns or data correlations, which in turn can be integrated with AI to improve results. Based on these characteristics, in a study of structured and unstructured databases, recurrent correlations are sought through the crossing of information until reaching a plausible solution, which generally allows anticipating an event, such as COVID- 19. That is, it can be outlined in various contexts such as the spread of the virus in a certain area under climatological parameters, demographic density factors, mobility patterns, phenotypic characteristics of the virus with respect to its other families and environment, detection of abnormalities at the cellular level and / or molecular, among

many other aspects. For this type of development, Big Data uses AI through various algorithms applied either in deep learning as in machine learning mentioned above.

3.5 ETHICAL ASPECTS OF USING MACHINE LEARNING

Although the work of AI with Big Data and other disruptive technologies that accompany them are making their significant contribution to the health sector, it cannot be ignored that this type technological development leaves a gap in bioethics and information security, because data of patients and victims of COVID-19 is manipulated in a massive way, whose clinical history and traceability on monitoring of people and communities in general underlies both public and private repositories and, although access to these is restricted in some cases, even with data stored and shared in encrypted form, it would be lent to third parties to use it in discriminatory studies, where anonymity does not exist, as indicated by Gené, Gallo and de Lecuona (2018, p. 4), since the anonymization of personal data does not sufficiently guarantee privacy, specifically because massive data technology allows the re-identification of people. Lecuona and Villalobos (2018, p. 2) affirm that, by assigning a person or group particular characteristics, an individual becomes a component of a group that raises concerns about conscious and unconscious discrimination as a result of the use of big data in decision making. Discrimination under this approach to disruptive technologies is a risk to take into account, especially when decision-making is increasingly automated. Therefore, it is advisable that both data scientists, engineers, the State and the health sector, among others involved, evaluate the risk-benefit of the use of these technologies, not restricting them, but creating the legal, technical and technological mechanisms that guarantee that no right is violated, and a first step in this direction is to ensure that clinical information is not outsourced without any type of restriction (De Lecuona, 2018). With the continuous incorporation of disruptive technologies in the field of health, regulations regarding the preservation of patient rights should change and adapting despite the uncertainty it brings, which as López (2019) points out, not only do we underestimate the effects of technology in the short term, but we are not even capable of forecasting, predicting or intuiting its long-term effects.

4. PROPOSED METHOD

4.1 MATHEMATICAL MODELLING

The population growth curves follow the well-known logistic behavior as shown in figure 1. A model that is used to fit the population curves is the Logistic one, which uses the following equation:

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

As you can see, this model has several problems including that the data is in $[0, 1]$, and it is not flexible, the advantage is that this model is that an approximation of the optimal solution can be obtained by transforming the data and using regression linear. As seen in graph 2, when applying logistic regression and dividing all the data by the maximum value (66818), a curve that fits well is obtained, with $R^2 = 0.99955$, which is very good, from a statistical point of view, but as seen in the figure there are many values that do not fit very well in the curves and it is not good for prediction, as we will see later.

The classic version for the study of epidemics is the SIR model in which the population is divided into three groups: the susceptible, the infected and the recovered (SIR), this in the simplest case and that the population is changing from susceptible to infected and then recovered. It is assumed that the decrease is proportional to the number of infected multiplied by the same number of susceptible. the change in the recovered is equal to a certain percentage of the infected and finally the number of infected will change increasing by the susceptible and then we remove the amount of the infected that passes recovered. In this way the following equations are obtained:

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0,$$

$$S(t) + I(t) + R(t) = \text{constant} = N,$$

The problem with the SIR model is that as we can see, the three parameters that are in the differential equations are needed but with little data or say in a few days it is very difficult to determine those parameters.

4.2 LOGISTIC REGRESSION

Therefore, the use of a model a little more complex than the logistic regression is proposed and that is easier to determine than the SIR. A first version is:

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

which adds one more parameter, to be determined, which is the limit population M , since with this modification it cannot be solved by transformation and linear regression. To solve this problem, non-linear optimization techniques must be used to solve it,

It also has the problem of rigidity, that is, it does not fit enough to certain curves, this model can be used for prediction, but from the point of adjustment it is not as accurate as we will see later. To make the curve more flexible, an extra parameter α is added as follows:

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

This α parameter adds flexibility in fitting, remember the graphs of $y = x^{1/3}$, $y = x^{1/2}$, $y = x$, $y = x^2$, $y = x^3$. For this function the inflection point is when:

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

so we have that the inflection point is obtained for:

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

This is the same model known as the Richards curve that is used to model population growth. Note that with some calculations and $A = 0$ equality is given with generalized logistics. The case:

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

which is a solution of the differential equation:

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

The proposal of this work is to first use the generalized logistic curve using MATLAB r2018a to make an adjustment of the data in which the curve is almost complete, for example data from China and South Korea, (30 of March) On the other hand, when all the data are available there is

a regression type curve Generalized logistics that fit the data. So, the hypothesis here is: If I have the lower part of the Curve, that is, the first values of the curve, I can obtain the parameters of the curve, and obtain the complete curve. And with this it can be used to predict population growth, in this case, for example, the total number of cases due to covid-19 in a country or region and when the inflection point is reached, that is, moment in time. that the number of daily cases begins to decrease.

More specifically, if you have data for 20 to 30 or 35 days, the question is: can you determine the parameters of the curve that fits the complete data? If this is achieved as we see, we would have a way of predicting the behavior of population growth with only a few days, that is, we can determine the parameters of the curve with a few days, which is very difficult with the SIR model to determine the parameters of the differential equations.

4.3 ADJUSTING THE CURVE

For this work we are using the data provided by the European Center for Disease Prevention and Control on the page to download daily data see [1]. The first thing that is presented is to verify, as is known, that the LG curve fits the covid-19 data very well. In the case of China and South Korea, which are the ones with the almost complete curve. Using a non-linear optimization algorithm, the following results are obtained.

4.3.1 China Data

With $R^2 = 0.998878$. As can be seen, a very good approximation is obtained, but as we know and can be seen in figure 3, there is a jump in these data, so it was decided to correct this jump by putting the data of 02/13/2020 equal to the day previous and the day 02/14/2020 equal to the following day. With these corrections, the following data is obtained and the graph in figure:

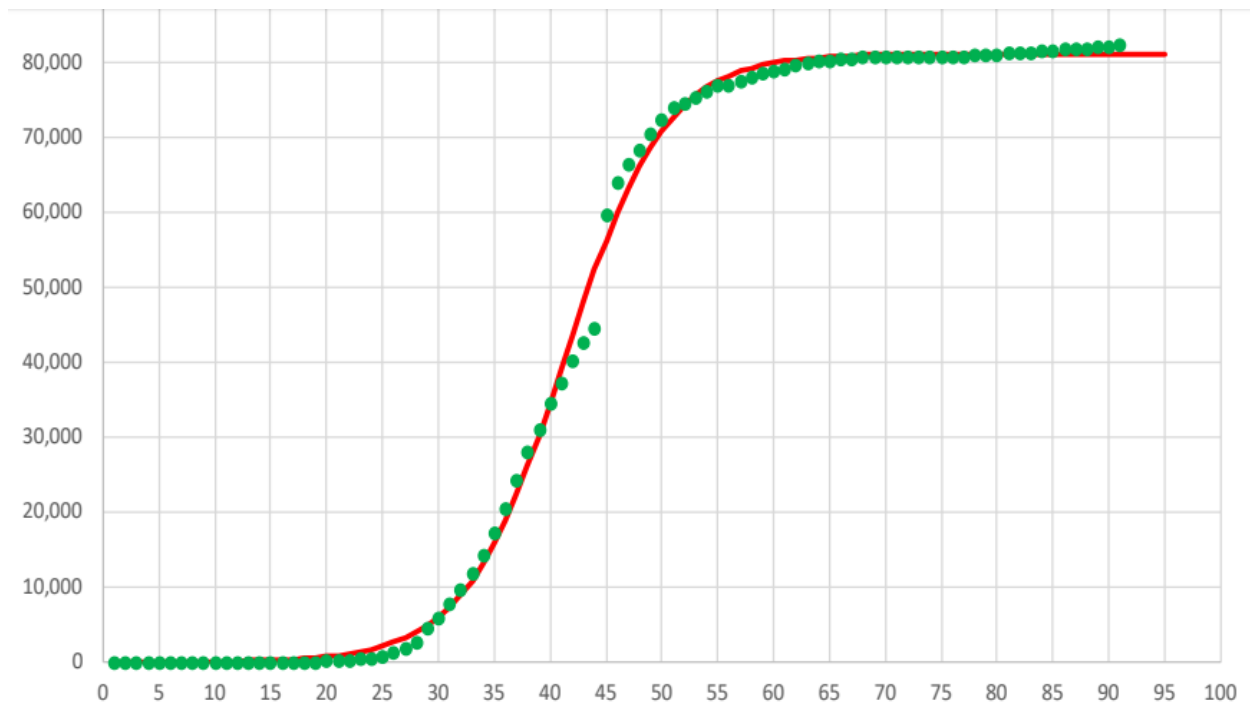


Figure 4.1: China data where the green dots are the actual cases and the red line is the predicted cases

In this case, an $R^2 = 0.999885$ is obtained, which is better than that of the data without correction. Note that in the latest data it is seen that there is a linear trend, this can be improved as seen in the case of South Korea, as seen in the next subsection.

4.3.2 South Korea Data

When applying the method to the data from South Korea, the following data is obtained and the graph in figure Y gives an $R^2 = 0.99875$, which is not very good, as seen in the figure that the fit is not very good:

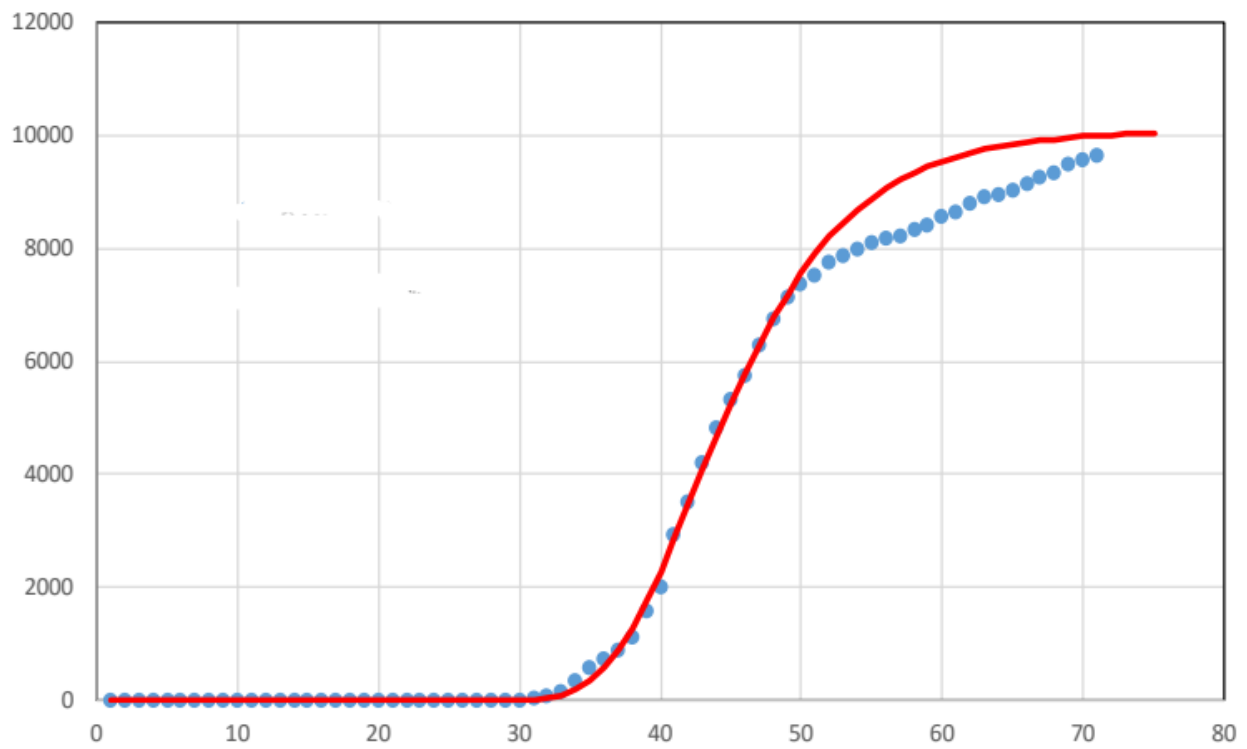


Figure 4.2: S-Korea data where the Blue dots are the actual cases and the red line is the predicted cases

5. SIMULATION AND RESULTS

As seen in the previous section, the logistic curves approximate very well, as has been shown in previous works (for example [2, 4]). In this section we are going to try to measure the percentage of relative error that is committed when using the proposed curves, a test curve and data from China and South Korea will be used and compared to previous studies and then applied to various countries which includes Turkey and Iraq.

5.1 TURKEY DATA

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 curve is almost complete. As we see when using the days from 16 to 45 (20 days), note that the first 16 days there were almost no changes in the data so that data was not used, with these data with LG you get a maximum error of 9.91% is obtained (average of 3 runs, basically gives the same value) When using the days from 15 to 55 (30 days) the graph of figure below is obtained, when measuring the relative errors a maximum value 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:

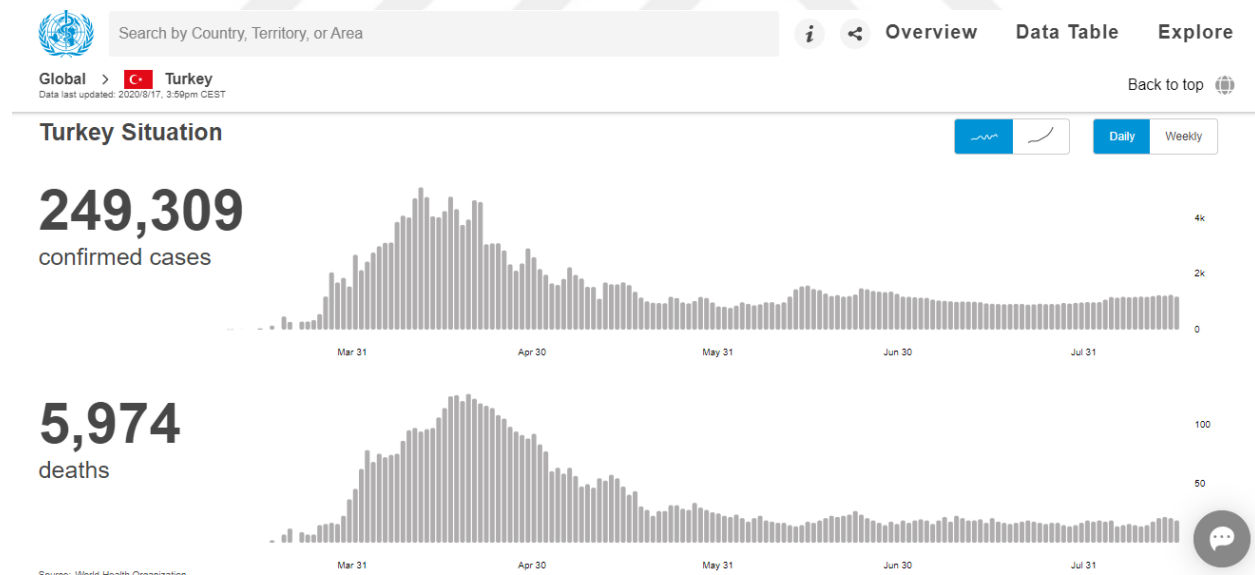


Figure 5.1: Covid-19 Cases in turkey from WHO website

5.2 IRAQ DATA

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:

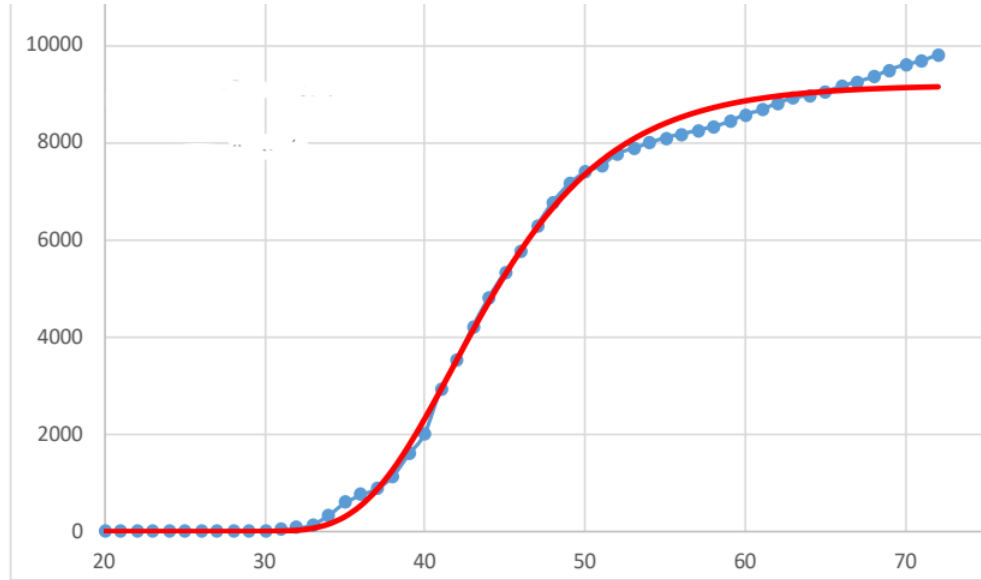


Figure 5.2: Data from IRAQ and (real and predicted)

The rest are simulation done on various countries on the data that we obtained from WHO website:

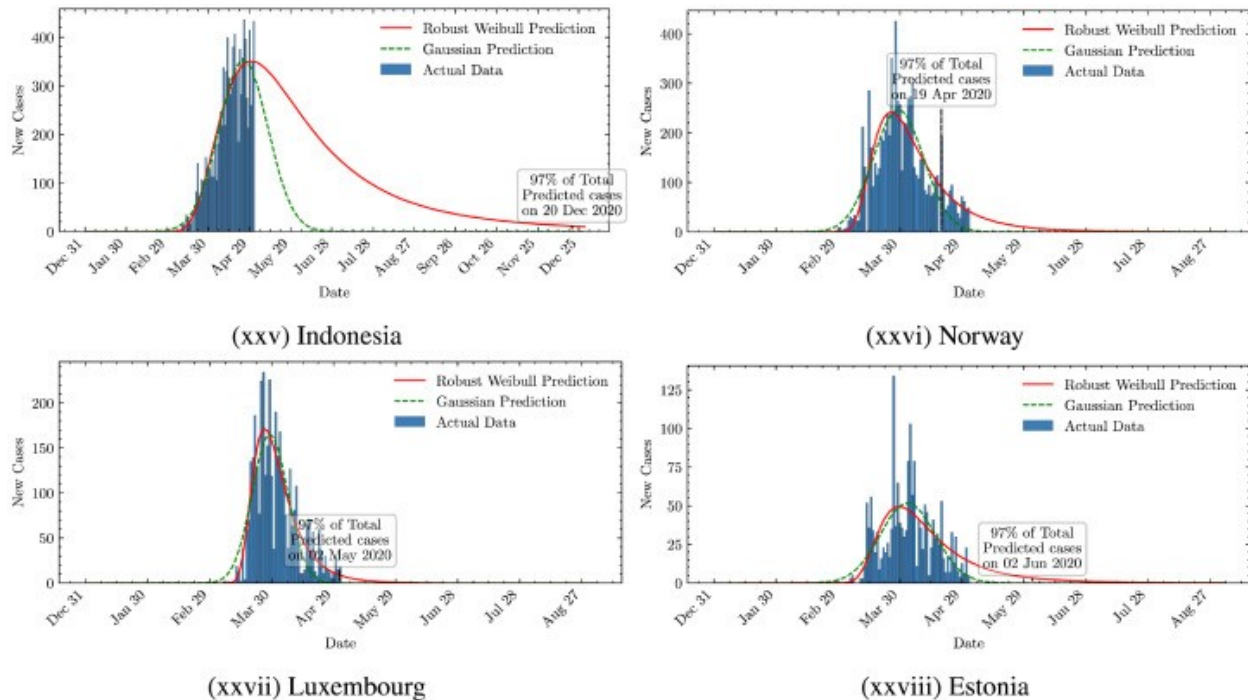


Figure 5.3: Population growth curves can be adjusted, using the LG 1

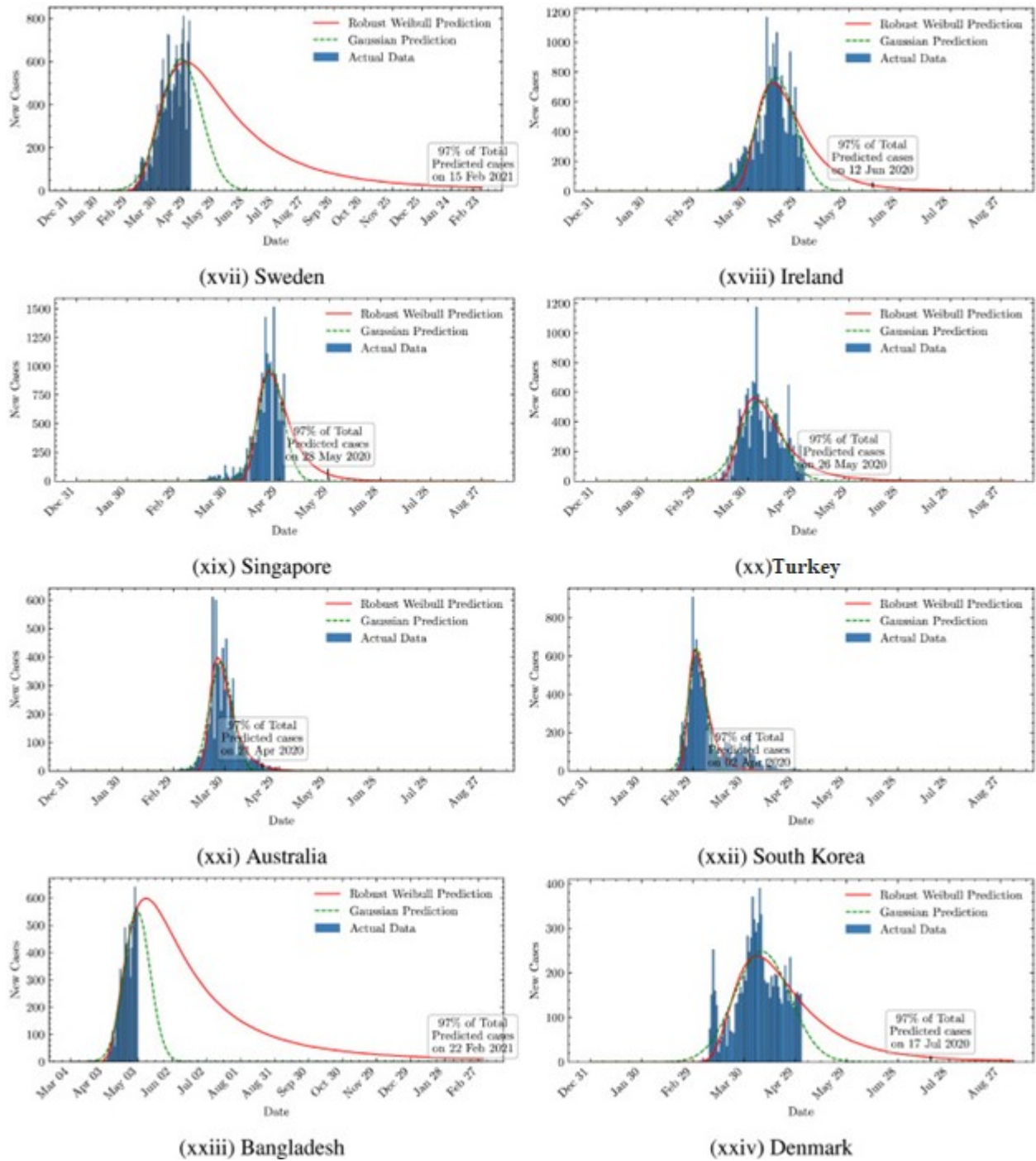


Figure 5.4: Population growth curves can be adjusted, using the LG 2

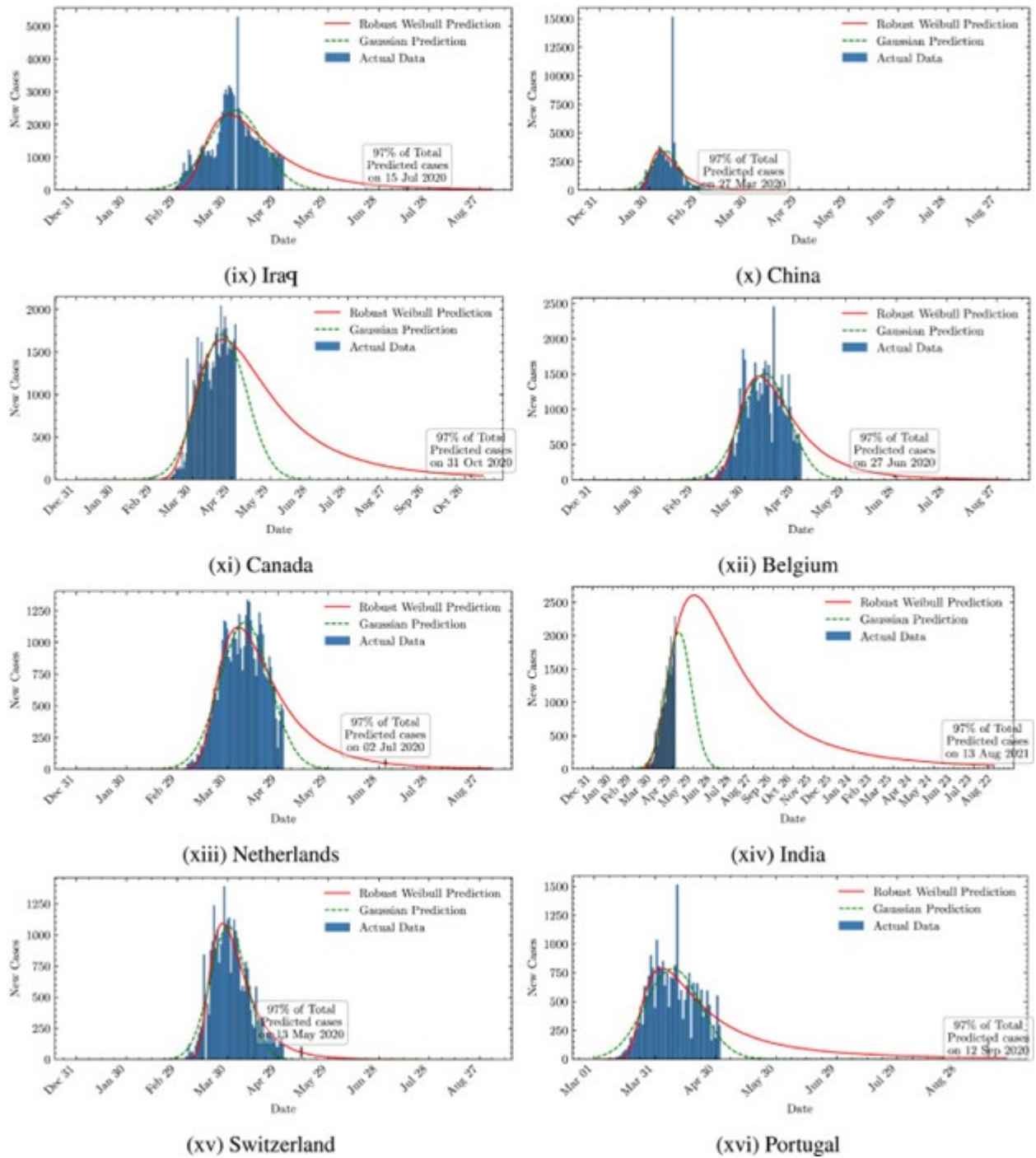


Figure 5.5: Population growth curves can be adjusted, using the LG 3

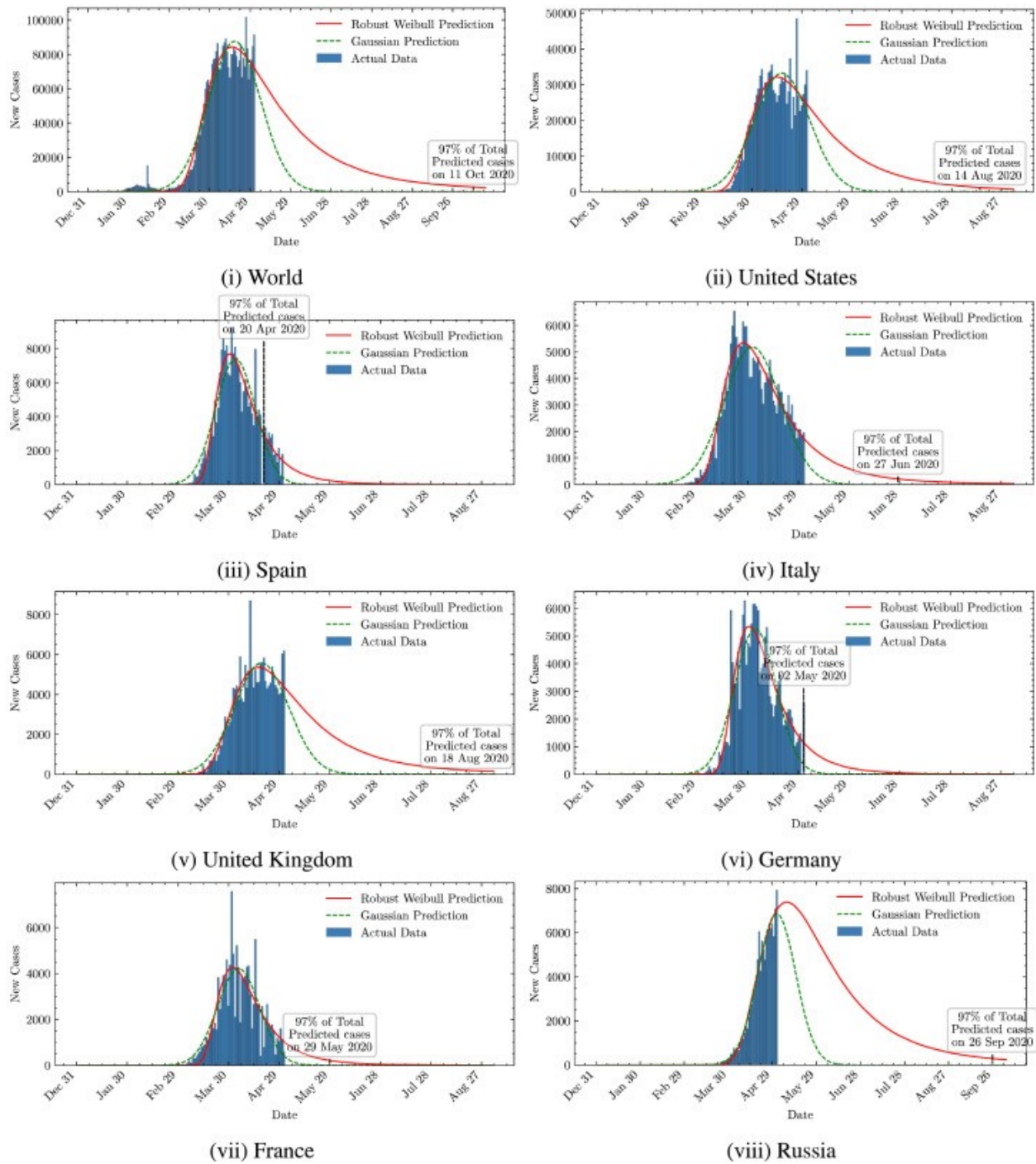


Figure 5.6: Population growth curves can be adjusted, using the LG 4

This work shows how population growth curves can be adjusted, using the LG and Gompers functions, even in the most general case as in the data for South Korea, China, Iraq, Turkey section 5.2, where logistics are being carried out with a line. It is seen that these methods could be used to predict population growth, in this case of people infected by Covid-19, and could help pandemic experts to take the necessary measures.

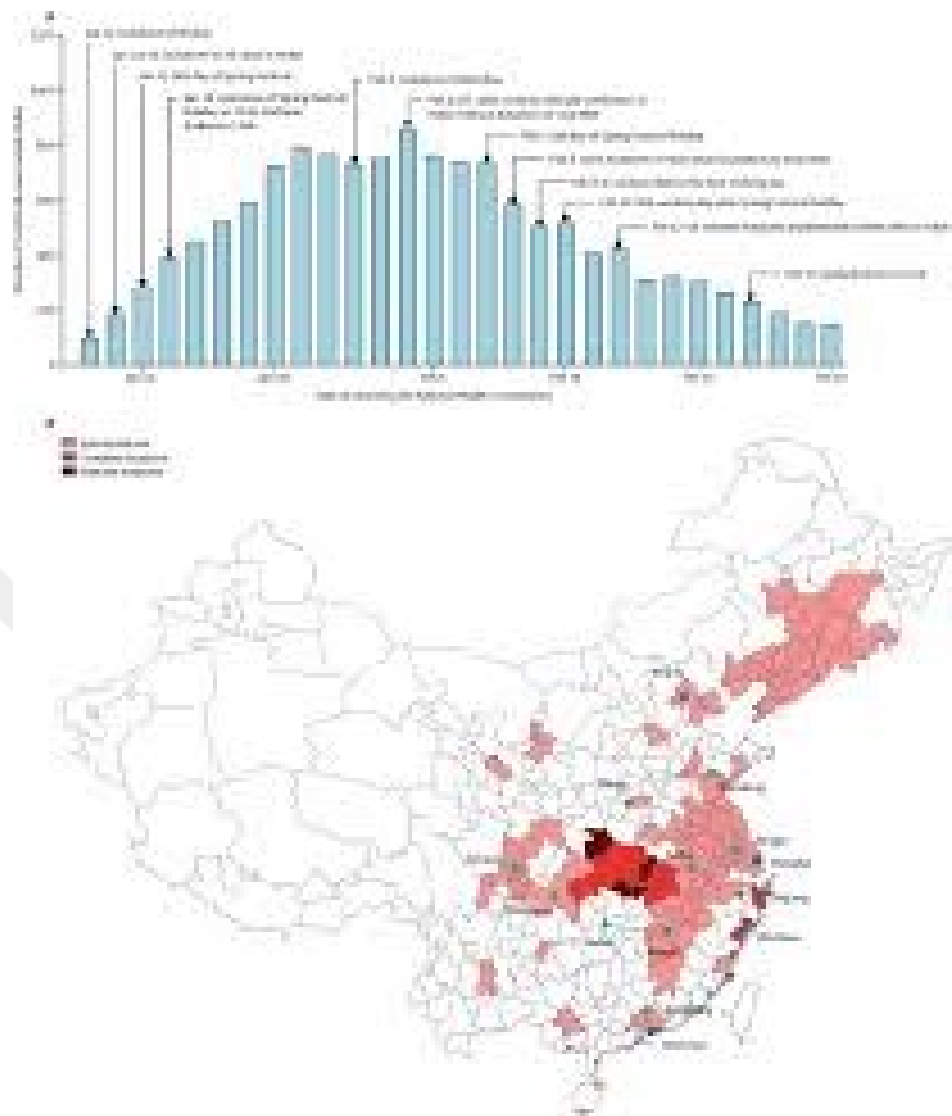


Figure 5.7: China population related to Covid-19 cases

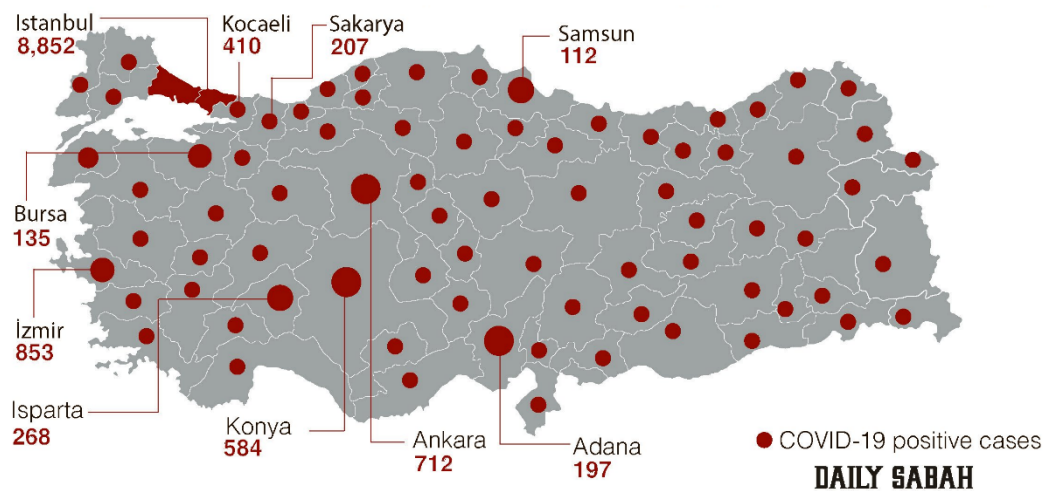
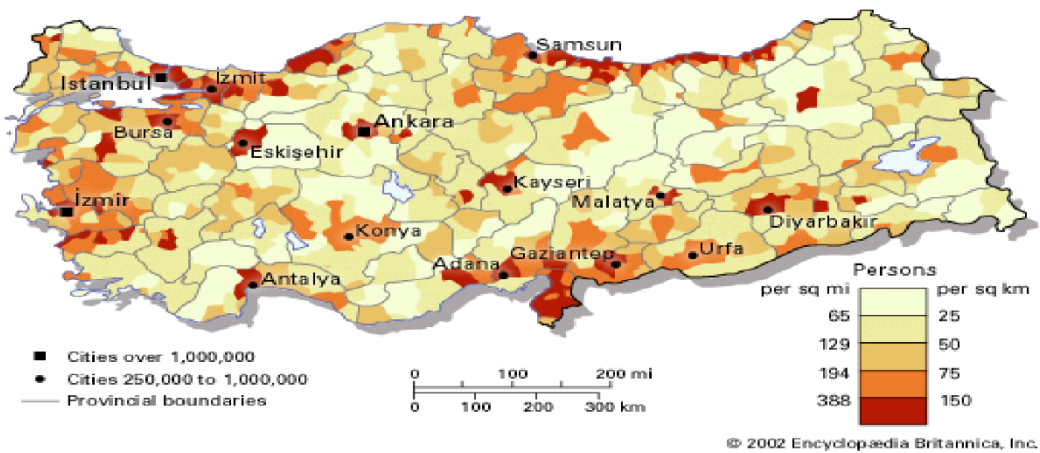


Figure 5.8: Turkey Covid-19 cases and population density map

6. 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 radio diagnosis 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.

REFERENCES

- [1] W. H. Organization, "Coronavirus disease 2019 (COVID-19): situation report," 2020.
- [2] "CDC Works 24/7," 2020. [Online]. Available: <https://www.cdc.gov/>. [Accessed April 25,2020].
- [3] "SARS CORONA Virus," [Online]. Available: <https://www.sinobiological.com/research/virus/sars-coronavirus-overview>. [Accessed April 25,2020].
- [4] Al-Turaiki, M. Alshahrani and T. Almutair, "Building predictive models for MERS-CoV infections using data mining technique," *Journal of Infection and Public Health* (2016), vol. 9, p. 744—748, 2016.
- [5] Du Toit, "Outbreak of a novel coronavirus," *Nature Reviews Microbiology*, vol. 18, no. 3, pp. 123-123, 2020
- [6] W. Wang, J. Tang and F. Wei, "Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China," *Journal of medical virology*, vol. 92, no. 4, pp. 441-447, 2020.
- [7] W. Naudé, "Artificial Intelligence against COVID-19: An Early Review," *IZA Discussion Papers* 13110, 2020
- [8] S. F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A. R. VarkonyiKoczy, U. Reuter, T. Rabczuk and P. M. Atkinson, "COVID-19 Outbreak Prediction with Machine Learning, 2020040311," *Preprints* 2020, doi: 10.20944/preprints202004. 0311.v1
- [9] S. Tuli, S. Tuli, R. Tuli and S. S. Gill, "Predicting the Growth and Trend of COVID-19 Pandemic using Machine Learning and Cloud Computing," *Internet of Things*, p. 100222, 2020.
- [10] P. N. Mahalle, N. P. Sable, N. P. Mahalle and G. R. Shinde, "Predictive Analytics of COVID-19 Using Information, Communication and Technologies," *Preprints* 2020, 2020040257 (doi: 10.20944/preprints202004. 0257.v1).

- [11] D. Yadav, H. Maheshwari and U. Chandra, "Outbreak prediction of covid-19 in most susceptible countries," *Global Journal of Environmental Science and Management*, vol. 6, no. 4, p. 2020.
- [12] S. Das, A. Dey, A. Pal and N. Roy, "Applications of Artificial Intelligence in Machine Learning: Review and Prospect," *International Journal of Computer Applications*, vol. 115, pp. 31-41, 2015.
- [13] Palmer, J. Rafael and E. Gervilla, *Data Mining: Machine Learning and Statistical Technique*, 2011.
- [14] Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Lulu, 1st edition, 2019.
- [15] J. F. and J. Fang, "Why Logistic Regression Analyses Are More Reliable Than Multiple Regression Analyses," *Journal of Business and Economics*, vol. 4, no. 7, pp. 620-633, 2013.
- [16] V. Jakkula, "Tutorial on support vector machine (SVM)," School EECS,, Washington State Univ, Washington, DC, USA, 2011.
- [17] L. Dencelin and R. Thenmoley, "Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures," *Biomedical Research (2016) Computational Life Sciences and Smarter Technological Advancement*, pp. 166-S173, 2016.
- [18] J. Han and M. Kamber, "Data mining: concepts and techniques," San Francisco, Morgan Kaufmann Publishers, 2001.
- [19] E. MATTHES, *Python crash course: a hands-on, project-based introduction to programming*, 2016.
- [20] DOS (2015), "Jordanian Department of Statistics. Retrieved 10 4, 2016, from," [Online]. Available: http://dos.gov.jo/dos_home_e/main/linkedhtml/jordan_no.htm. [Accessed April 2020].

- [21] Chen Z, Zhang W, Lu Y, Guo C, Guo Z, Liao C, et al. From SARS-CoV to Wuhan 2019-nCoV Outbreak: Similarity of Early Epidemic and Prediction of Future Trends. *BioRxiv* [Preprint]. 2020 bioRxiv 919241 [posted 2020 January 27; cited 2020 January 31]. Available from: <https://www.biorxiv.org/content/10.1101/2020.01.24.919241v3>.
- [22] Hayward JA, Tachedjian M, Cui J, Field H, Holmes EC, Wang L, et al. Identification of diverse full-length endogenous betaretroviruses in megabats and microbats. *Retrovirology*. 2013; 10. <https://doi.org/10.1186/1742-4690-10-35> PMID: 23537098
- [23] Cui J, Tachedjian G, Wang LF. Bats and Rodents Shape Mammalian Retroviral Phylogeny. *Sci. Rep.* 2015; 5. <https://doi.org/10.1038/srep16561>
- [24] Hayward JA, Tachedjian M, Cui J, Cheng AZ, Johnson A, Baker ML, et al. Differential evolution of antiretroviral restriction factors in pteropid bats as revealed by APOBEC3 gene complexity. *Mol. Biol. Evol.* 2018; 35: 1626–1637. <https://doi.org/10.1093/molbev/msy048> PMID: 29617834
- [25] Wong A, Li X, Lau S, Woo P. Global Epidemiology of Bat Coronaviruses. *Viruses*. 2019; 11(2): 174. <https://doi.org/10.3390/v11020174>
- [26] Yang XL, Hu B, Wang B, Wang MN, Zhang Q, Zhang W, et al. Isolation and Characterization of a Novel Bat Coronavirus Closely Related to the Direct Progenitor of Severe Acute Respiratory Syndrome Coronavirus. *J. Virol.* 2016; 90: 3253–3256. <https://doi.org/10.1128/JVI.02582-15>
- [27] Lau SK, Li KS, Tsang AK, Lam CS, Ahmed S, Chen H, et al. Genetic Characterization of Betacoronavirus Lineage C Viruses in Bats Reveals Marked Sequence Divergence in the Spike Protein of Pipistrellus Bat Coronavirus HKU5 in Japanese Pipistrelle: Implications for the Origin of the Novel Middle East Respiratory Syndrome Coronavirus. *J. Virol.* 2013; 87: 8638–8650. <https://doi.org/10.1128/JVI.01055-13> PMID: 23720729

- [28] Lacroix A, Duong V, Hul V, San S, Davun H, Omaliss K, et al. Genetic diversity of coronaviruses in bats in Lao PDR and Cambodia. *Infect. Genet. Evol.* 2017; 48: 10–18. <https://doi.org/10.1016/j.meegid.2016.11.029> PMID: 27932284
- [29] Drexler JF, Gloza-Rausch F, Glende J, Corman VM, Muth D, Goettsche M, et al. Genomic Characterization of Severe Acute Respiratory Syndrome-Related Coronavirus in European Bats and Classification of Coronaviruses Based on Partial RNA-Dependent RNA Polymerase Gene Sequences. *J. Virol.* 2010; 84: 11336–11349. <https://doi.org/10.1128/JVI.00650-10> PMID: 20686038
- [30] Rihtarič D, Hostnik P, Steyer A, Grom J, Toplak I. Identification of SARS-like coronaviruses in horseshoe bats (*Rhinolophus hipposideros*) in Slovenia. *Arch. Virol.* 2010; 155: 507–514. <https://doi.org/10.1007/s00705-010-0612-5> PMID: 20217155
- [31] He B, Zhang Y, Xu L, Yang W, Yang F, Feng Yun, et al. Identification of Diverse Alphacoronaviruses and Genomic Characterization of a Novel Severe Acute Respiratory Syndrome-Like Coronavirus from Bats in China. *J. Virol.* 2014; 88: 7070–7082. <https://doi.org/10.1128/JVI.00631-14> PMID: 24719429
- [32] Wacharapluesadee S, Duengkae P, Rodpan A, Kaewpom T, Maneeorn P, Kanchanasaka B, et al. Diversity of coronavirus in bats from Eastern Thailand Emerging viruses. *Virol. J.* 2015; 12: 1–7. <https://doi.org/10.1186/s12985-015-0289-1>
- [33] Tong S, Conrardy C, Ruone S, Kuzmin IV, Guo X, Tao Y, et al. Detection of novel SARS-like and other coronaviruses in bats from Kenya. *Emerg. Infect. Dis.* 2009; 15: 482–485. <https://doi.org/10.3201/eid1503.081013> PMID: 19239771
- [34] Lau SKP, Woo PCY, Li KSM, Huang Y, Tsoi H, Wong BHL, et al. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc. Natl. Acad. Sci.* 2005; National Academy of Sciences, U. S. A., 102: 14040–14045. <https://doi.org/10.1073/pnas.0506735102> PMID: 16169905
- [35] Virologists weigh in on novel coronavirus in China's outbreak. 2020 January 08 [cited 31 January 2020]. In: University of Minnesota [Internet]. Available from:

<http://www.cidrap.umn.edu/news-perspective/2020/01/virologists-weigh-novel-coronavirus-chinas-outbreak>.

- [36] nCoV's relationship to bat coronaviruses & recombination signals (no snakes)—no evidence the 2019- nCoV lineage is recombinant. 2020 January 31 [cited 31 January 2020]. In: Virological blog [Internet]. Available from: <http://virological.org/t/ncovs-relationship-to-bat-coronaviruses-recombination-signalsno-snakes-no-evidence-the-2019-ncov-lineage-is-recombinant/331>.
- [37] Experts: nCoV spread in China's cities could trigger global epidemic. 2020 January 27 [cited 31 January 2020]. In: University of Minnesota [Internet]. Available from: <http://www.cidrap.umn.edu/newsperspective/2020/01/experts-ncov-spread-chinas-cities-could-trigger-global-epidemic>.
- [38] China detects large quantity of novel coronavirus at Wuhan seafood market. 2020 January 27 [cited 31 January 2020]. In: Xinhuanet News [Internet]. Available from: http://www.xinhuanet.com/english/2020-01/27/c_138735677.
- [39] Zhao Y, Zhao Z, Wang Y, Zhou Y, Ma Y, Zuo W. Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCoV. *BioRxiv* [Preprint]. 2020 bioRxiv 919985 [posted 2020 January 26; cited 2020 January 31]. Available from: <https://www.biorxiv.org/content/10.1101/2020.01.26.919985v1>.
- [40] Li Y, Zhang J, Wang N, Li H, Shi Y, Gui G, et al. Therapeutic Drugs Targeting 2019-nCoV Main Protease by High-Throughput Screening. *BioRxiv* [Preprint]. 2020 bioRxiv 922922 [posted 2020 January 30; cited 2020 January 31]. Available from: <https://www.biorxiv.org/content/10.1101/2020.01.28.922922v2>.
- [41] Liu X, Wang XJ. Potential inhibitors against 2019-nCoV coronavirus M protease from clinically approved medicines. *Journal of Genetics and Genomics*. 2020; 47(2): 119–121. <https://doi.org/10.1016/j.jgg.2020.02.001> PMID: 32173287
- [42] Vingia S, Almeida J. Alignment-free sequence comparison—a review. *Bioinformatics*. 2003; 19(4): 513– 523. <https://doi.org/10.1093/bioinformatics/btg005> PMID: 12611807

- [43] Zielezinski A, Vinga S, Almeida J, Karlowski WM. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biology*. 2017, 18: 186. <https://doi.org/10.1186/s13059-017-1319-7> PMID: 28974235
- [44] Kari L, Hill KA, Sayem AS, Karamichalis R, Bryans N, Davis K, et al. Mapping the space of genomic signatures. *PLoS ONE*. 2015; 10: e0119815. <https://doi.org/10.1371/journal.pone.0119815> PMID: 26000734
- [45] Karamichalis R, Kari L, Konstantinidis S, Kopecki S. An investigation into inter- and intragenomic variations of graphic genomic signatures. *BMC Bioinformatics*. 2015; 16: 246. <https://doi.org/10.1186/s12859-015-0655-4> PMID: 26249837
- [46] Solis-Reyes S, Avino M, Poon A. An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PLoS ONE*. 2018; 13: e0206409. <https://doi.org/10.1371/journal.pone.0206409> PMID: 30427878
- [47] Randhawa GS, Hill KH, Kari L. ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. *BMC Genomics*. 2019; 20: 267. <https://doi.org/10.1186/s12864-019-5571-y> PMID: 30943897
- [48] Randhawa GS, Hill KH, Kari L. MLDSP-GUI: an alignment-free standalone tool with an interactive graphical user interface for DNA sequence comparison and analysis. *Bioinformatics*. 2019; btz918.
- [49] Jeffrey HJ. Chaos game representation of gene structure. *Nucleic Acids Res*. 1990; 18: 2163–2170. <https://doi.org/10.1093/nar/18.8.2163> PMID: 2336393
- [50] Asuero AG, Sayago A, González AG. The correlation coefficient: an overview. *Crit Rev Anal Chem*. 2006; 36(1): 41–59. <https://doi.org/10.1080/10408340500526766>
- [51] Karamichalis R, Kari L. MoDMaps3D: an interactive webtool for the quantification and 3D visualization of interrelationships in a dataset of DNA sequences. *Bioinformatics*. 2017; 33(19): 3091–3. <https://doi.org/10.1093/bioinformatics/btx367> PMID: 28605460

- [52] Kruskal J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*. 1964; 29: 1–27. <https://doi.org/10.1007/BF02289565>
- [53] Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*. 1958; 38: 1409–1438.
- [54] Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*. 1987; 4(4): 406–425. <https://doi.org/10.1093/oxfordjournals.molbev.a040454> PMID: 3447015
- [55] Carneiro RL, Requião RD, Rossetto S, Domitrovic T, Palhano FL. Codon stabilization coefficient as a metric to gain insights into mRNA stability and codon bias and their relationships with translation. *Nucleic acids research*. 2019; 47(5): 2216–2228. <https://doi.org/10.1093/nar/gkz033> PMID: 30698781
- [56] Karumathil S, Raveendran NT, Ganesh D, Kumar NS, Nair RR, Dirisala VR. Evolution of Synonymous Codon Usage Bias in West African and Central African Strains of Monkeypox Virus. *Evolutionary Bioinformatics Online*. 2018; 14. <https://doi.org/10.1177/1176934318761368> PMID: 29551886
- [57] Vinogradov AE, Anatskaya OV. DNA helix: the importance of being AT-rich. *Mammalian Genome*. 2017; 9(10): 455–464. <https://doi.org/10.1007/s00335-017-9713-8>
- [58] Hollander M, Wolfe DA, Chicken E. *Nonparametric statistical methods*, 3rd Edition, John Wiley & Sons; 2013.
- [59] Zhao S, Lin Q, Ran J, Musa SS, Yang G, Wang W, et al. Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *International Journal of Infectious Diseases*. 2020; 92: 214–217. <https://doi.org/10.1016/j.ijid.2020.01.050> PMID: 32007643
- [60] Shao P, Shan Y. Beware of asymptomatic transmission: Study on 2019-nCoV prevention and control measures based on extended SEIR model. *BioRxiv [Preprint]*. 2020 bioRxiv 923169

APPENDIX A.

A.1 ALL THE CODES USED IN MATLAB R2018A

```
%author
%Astick Banerjee
%Disclaimer: since it is Spline Interpolation the expected value may
not
%be correct

%Covid-19 China Unofficial json api of official data
history_api= 'https://api.rootnet.in/covid19-in/stats/history';
char expectedValue;
%covidPatientDataIndia = webread(api);
covidHistoryData = webread(history_api);
data = covidHistoryData.data;
n = length(data);
totalArray = [];
for i =1:n
    date = data(i).day;
    total = data(i).summary.total;
    totalArray = [totalArray,total];
    %disp(date)
    %disp(totalArray)
end
%Exact data of covid-19 India
today = data(n).day;
%disp(today)
dayArray = 1:1:n;
plot(dayArray,totalArray,'bo'); hold on;

%Interpolation using Spline
% n+2 denotes the next two day prediction
dayA = 1:1:n+1;
sArr = spline(dayArray,totalArray,dayA);
plot(dayA,sArr,'-k');hold on;
text(dayArray(n)+.3,sArr(n),['(' today ' , ' num2str(sArr(n)),')'])
plot(dayA(n+1),sArr(n+1),'ro');
title('Covid - 19 Case Prediction')
%tday = split(today,"-");
%u =str2double(tday(3))+3;
%disp(u)
text(dayA(n+1)+.3,sArr(n+1),['( Tomorrow , '
num2str(int32(sArr(n+1))),')'])
xlabel('Days')
ylabel('Total Case')
legend({'Original', 'Spline Interpolation',
'Expected'},'Location','northwest' );
expectedValue = int32(sArr(n+1));
disp("Today's total Covid-19 +ve Case : " + int32(sArr(n)))
disp("Tomorrow will be (expected): " + expectedValue)

imds = imageDatastore('dataset', ...
    'IncludeSubfolders',true, ...
    'LabelSource','foldernames');
```

```

[imdsTrain,imdsTest] = splitEachLabel(imds, 0.70,'randomize');

augimdsTrain = augmentedImageDatastore([224 224], imdsTrain,
'ColorPreprocessing', 'gray2rgb');
augimdsTest = augmentedImageDatastore([224 224], imdsTest,
'ColorPreprocessing', 'gray2rgb');

%specify the layers in the network
layers = [
    imageInputLayer([224 224 3])

    convolution2dLayer(3,16,'Padding',1)
    batchNormalizationLayer
    reluLayer
    maxPooling2dLayer(2,'Stride',2)

    convolution2dLayer(3,32,'Padding',1)
    batchNormalizationLayer
    reluLayer
    maxPooling2dLayer(2, 'stride', 2)

    fullyConnectedLayer(2)
    softmaxLayer
    classificationLayer];

options = trainingOptions('sgdm', ...
    'MiniBatchSize',32, ...
    'MaxEpochs',4, ...
    'InitialLearnRate',3e-4, ...
    'Shuffle','every-epoch', ...
    'ValidationData',augimdsTest, ...
    'Verbose',false, ...
    'Plots','training-progress');

net = trainNetwork(augimdsTrain,layers,options);

[YPred,probs] = classify(net,augimdsTest);
accuracy = mean(YPred == imdsTest.Labels);

idx = randperm(numel(augimdsTest.Files),4);
figure
for i = 1:4
    subplot(2,2,i)
    I = readimage(imdsTest,idx(i));
    imshow(I)
    label = YPred(idx(i));
    title(string(label) + ", " + num2str(100*max(probs(idx(i),:)),3) +
"%");
end

```

```

close all;
clear all;
Img =image;
msk=image;
Img =double(Img(:,:,1));
A=255;
sigma = 4;
G=fspecial('gaussian',15,sigma);
Img=conv2(Img,G,'same');
nu=0.001*A^2; % coefficient of arc length term
sigma = 4; % scale parameter that specifies the size of the
neighborhood
iter_outer=50;
iter_inner=10; % inner iteration for level set evolution
timestep=.1;
mu=1; % cient for distance regularization term (regularize the level
set function)
c0=1;
figure(1);
imagesc(Img,[0, 255]); colormap(gray); axis off; axis equal
% initialize level set function
initialLSF = c0*ones(size(Img));
%initialLSF(30:90,50:90) = -c0;
%initialLSF(30:200,50:220) = -c0;
initialLSF(25:220,25:220) = -c0;
u=initialLSF;
hold on;
contour(u,[0 0],'r');
title('Initial contour');
figure(2);
imagesc(Img,[0, 255]); colormap(gray); axis off; axis equal
hold on;
contour(u,[0 0],'r');
title('Initial contour');
epsilon=1;
b=ones(size(Img)); %% initialize bias field
K=fspecial('gaussian',round(2*sigma)*2+1,sigma); % Gaussian kernel
KI=conv2(Img,K,'same');
KONE=conv2(ones(size(Img)),K,'same');
[row,col]=size(Img);
N=row*col;
for n=1:iter_outer
    [u, b, C]= lse_bfe(u,Img, b, K,KONE, nu,timestep,mu,epsilon,
iter_inner);

    if mod(n,2)==0
        pause(0.001);
        imagesc(Img,[0, 255]); colormap(gray); axis off; axis equal;
        hold on;
        contour(u,[0 0],'r');
        iterNum=[num2str(n), ' iterations'];
        title(iterNum);
        hold off;
    end
end
Mask =(Img>10);
Img_corrected=normalize01(Mask.*Img./(b+(b==0)))*255;
I=u
maxI = max(I(:));

```

```

for y=1:40
    fil(x,y)=0;
end
end
%%%%%%%%%%%%%%
for x=1:256
for y=230:256
    fil(x,y)=0;
end
end
figure(3);
imagesc(fil); colormap(gray); title('Final Binary Mask'); ax

```