Faculty of Graduate Studies
Master of Applied Statistics and Data Science

Predicting the Likelihood of Infection and Severity of Symptoms of COVID-19 in Palestine

توقع احتمالية الإصابة بفايروس كورونا وشدة الأعراض في فلسطين

Prepared By

Asmaa Mohamed Shuibi

1205039

Supervisors

Dr. Radi Jarrar and Dr. Sameera Awawda

Submitted in Partial Fulfillment of the Requirements of the "Master Degree in Applied Statistic and Data Science" from the Faculty of Graduate Studies at Birzeit University / Palestine

2023

جامعة بيرزيت
BIRZEIT UNIVERSITY

Predicting the Likelihood of Infection and Severity of Symptoms of COVID-19 in Palestine

Prepared By
Asmaa Mohamed Shuibi

Student Number: 1205039

Committee:

Dr. Radi Jarrar

Dr. Sameera Awada

Dr. Mohammad Abu-Zaineh

Dr. Anas Arram

Submitted in Partial Fulfillment of the Requirements for the "Master Degree in Applied Statistics and Data Science" from the Faculty of Graduate Studies at Birzeit University, Palestine.

2023

الإهـــداء

أهدي ثمرة بحثي هذا الى روح والدي الطاهرة وروح عمي ميسرة واخ زوجي نسيم....رحمهم الله جميعا واسكنهم فسيح جنانه

إلى من وقف إلى جانبي ودعمني ووهبه الله الصبر على كل ما عانيته ...زوجي ورفيقي

"عنان ميسرة الشعيبي"

لا تسعني الكلمات لأعبر عن كم الجهد الذي بذلته معي ... أدامك الله ذخرا لي... وسندا

إلى من ساندوني صبرا... في كل امتحان... وكل ورقة كتبتها اطفالي "نسيم، نادية، آسر، إيوان"

إلى "أمي الغالية... أطال الله عمرك ورعاك"

"اخوتي...سندي"

"أهل زوجي... القوة الداعمة في كل ظرف"

إلى كل النفوس السمحة الطيبة التي لا تحمل للناس إلا الخير و تضمر لهم غير المودة...

إلى كل الذين نحبهم ونشعر بأننا بهم أغنى الناس ... الى كل من علمنا حرفا وبالأخص منارات النور ومناهل العلم أساتذتي الأفاضل ...

إلى هؤلاء جميعا أهدي هذا البحث المتواضع ...

Acknowledgments

My heartfelt thanks go out to my advisors, Dr. Radi Jarrar and Dr. Sameera Awawda, for their invaluable guidance and support throughout my research and thesis writing journey. Their expertise and experience have been instrumental in helping me complete this project successfully.

I am deeply grateful to my family and friends who provided me with unwavering support and encouragement during my research. Their encouragement and motivation helped me stay on track and complete the project within the allotted time.

Special thanks for Dr. Mai Kaila and Dr. Leila Ghannam to provide all the facilities to complete this scientific thesis, and all those who helped me in the computer department in the Palestinian Health represented by Dr. Wael AL-sheikh, Dr. Kamal AL-Shakhra, Dr. Ali El-Helou, Dr. Samer Asaad, Dr. Randa Abu Rabi Responsible for the Corona-19 file at the World Health Organization in Palestine and all the staff... Thank you all.

Table of Contents

List of Figures

List of Tables

Abstract

The COVID-19 pandemic has had a global significant impact on various aspects of societies and resulted in a wide range of policy implications. Governments have implemented various measures, including lockdowns, mask mandates, and financial support for businesses and individuals. Moreover, various organizations and researchers tried to develop models to predict its spread. The main aim of these predictions is to help decision-makers take appropriate actions to slow the spread of the virus, reduce the number of cases, and mitigate the impact of the pandemic on the society and on the economy.

This thesis focuses on using Artificial Intelligence (AI) techniques, in particular, Machine Learning (ML) algorithms to analyze and study factors affecting the likelihood of contracting COVID-19 as well as the severity of its symptoms in Palestine. The study will use various ML algorithms to determine the best model for predicting both the likelihood of infection and the severity of symptoms. The main dataset used in this study is provided by the Palestinian Ministry of Health (PMoH). We also used records of hospitalized COVID-19 patients from hospitals located in different governorates in the West Bank.

The main advantage of the proposed approach is its ability to save time and increase the accuracy of detecting the likelihood of contracting COVID-19 and the severity of its symptoms while using large sets of data. Various models have been created and compared

The models applied show that several variables, including gender, result date, test type, cause of test, age, and age squared, are statistically significant in relation to the probability of contracting COVID-19. The results indicate that males tend to have a higher probability of contracting COVID-19 than females. Waves of the pandemic and certain regions and test types also affect the probability of contracting COVID-19.

Moreover, we also found that gender, result date, test type, cause of test, and age are statistically significant variables in determining the likelihood of contracting COVID-19. Males have a higher probability of contracting COVID-19 compared to females. The probability of contracting COVID-19 was higher during specific waves of the pandemic. Certain regions and test types also affect the probability of contracting COVID-19.

Additionally, the analysis revealed that individuals living in certain areas, such as Ramallah, Jenin, Jericho, Tubas, and Salfit have a higher probability of contracting COVID-19 than those living in Gaza. PCR testing shows a higher likelihood of COVID-19 infection compared to the AG test. Urban areas have a higher probability of contracting COVID-19 than refugee camps. As well, individuals with different reasons for testing and workers or travelers have a higher probability of contracting COVID-19.

The analysis revealed several significant factors influencing the severity of COVID-19 symptoms. These factors include the hospital where the patient is being treated, the department of treatment, eosinophil levels, test type, presence of blood diseases, and type of drugs. Patients treated at Yatta hospital have a higher probability of severe symptoms compared to Beit Jala Hospital. Treatment in the COVID-ICU department is associated with a higher probability of severe symptoms compared to the Intensive Care Department or Cardiac Intensive Care Department. High eosinophil levels, both test types, and the presence of blood diseases are also linked to a higher probability of severe symptoms. Furthermore, individuals who have not been vaccinated have a higher likelihood of experiencing severe symptoms. The best model for subjective severity is Model 3, while Model 6 provides the most accurate representation for objective severity.

The study found that most of the models accurately predict the likelihood and severity, with values ranging between 80% - 99%, which indicates the strength and accuracy of the models.

الملخص

انتشر فيروس كوفيد-19 عالميًا وأثر بشكل كبير على مختلف جوانب المجتمعات، وأدى إلى تبعات سياسية واسعة النطاق. فقد قامت الحكومات بتنفيذ تدابير مختلفة، بما في ذلك حظر التجوال وإلزامية ارتداء الكمامات وتقديم الدعم المالي للأعمال والأفراد. علاوة على ذلك، حاولت مختلف المنظمات والباحثون تطوير نماذج للتنبؤ بانتشار الفيروس. ويهدف هذا البحث إلى استخدام تقنيات الذكاء الاصطناعي، وبالأخص خوارزميات التعلم الآلي، لتحليل ودراسة العوامل التي تؤثر في احتمالية الإصابة بكوفيد-19 وشدة أعراضه في فلسطين. سيتم استخدام خوارزميات التعلم الآلي المختلفة لتحديد أفضل نموذج للتنبؤ بكل من احتمالية الإصابة وشدة الأعراض. وستستخدم مجموعة البيانات الأساسية في هذه الدراسة التي تم توفيرها من قبل وزارة الصحة الفلسطينية. تم استخدام سجلات المرضى المصابين بكوفيد-19 في المستشفيات الموجودة في محافظات مختلفة في الضفة الغربية.

الميزة الرئيسية للنهج المقترح هي قدرته على توفير الوقت وزيادة دقة اكتشاف احتمالية الإصابة بكوفيد-19 وشدة أعراضه باستخدام مجموعات كبيرة من البيانات. في هذا العمل، تم إنشاء العديد من النماذج ومقارنتها.

أظهرت النماذج المطبقة أن هناك عدة متغيرات ذات أهمية إحصائية تؤثر في احتمالية الإصابة بكوفيد-19. أشارت النتائج إلى أن الذكور لديهم احتمالية أعلى للإصابة بكوفيد-19 مقارنة بالإناث. تؤثر الموجات المتعاقبة للجائحة وبعض المناطق وأنواع الاختبار أيضًا على احتمالية الإصابة بكوفيد-19.

النتائج تشير إلى أن الجنس وتاريخ النتيجة ونوع الاختبار وسبب الاختبار والعمر هي متغيرات ذات أهمية إحصائية في تحديد احتمالية الإصابة بكوفيد-19. فالذكور لديهم احتمالية أعلى للإصابة بكوفيد-19 مقارنة بالإناث. كانت احتمالية الإصابة بكوفيد-19 أعلى خلال موجات محددة من الجائحة. تؤثر بعض المناطق وأنواع الاختبار أيضًا على احتمالية الإصابة بكوفيد-19.

بالإضافة إلى ذلك، يوجد احتمالية أعلى للإصابة بكوفيد-19 للأفراد الذين يعيشون في مناطق معينة مثل رام الله وجنين وأريحا وطوباس وسلفيت بالمقارنة بأولئك الذين يعيشون في قطاع غزة. يظهر الاختبار بي سي آر احتمالية أعلى للإصابة بكوفيد-19 مقارنة بالاختبار بالمضادات الجينية. تظهر المناطق الحضرية احتمالية أعلى للإصابة بكوفيد-19 مقارنة بمخيمات اللاجئين. علاوة على ذلك، يوجد احتمالية أعلى للإصابة بكوفيد-19 للأفراد الذين يخضعون لاختبار لأسباب مختلفة وللعاملين أو المسافرين.

كشف التحليل عن عدة عوامل مهمة تؤثر في شدة أعراض كوفيد-19. تشمل هذه العوامل المستشفى الذي يتم فيه علاج المريض، وقسم العلاج، ومستويات الإيوزينوفيل، ونوع الاختبار، ووجود أمراض الدم، ونوع الأدوية. المرضى الذين يتلقون العلاج في مستشفى يطا لديهم احتمالية أعلى لظهور أعراض شديدة مقارنة بمستشفى بيت جالا. قد يكون للإيوزينوفيل دور في تحديد شدة الأعراض أيضًا.

بشكل عام، يوضح هذا البحث قدرة تقنيات الذكاء الاصطناعي وخوارزميات التعلم الآلي على تحليل وتوقع احتمالية الإصابة بكوفيد-19 وشدة أعراضه في فلسطين. ومن المهم استمرار البحث وتحسين النماذج لتعزيز قدرتها التنبؤية وفهم العوامل المؤثرة في انتشار الفيروس وتفاعله مع السكان.

Acronyms

AI: Artificial Intelligence

ML: Machine Learning

ICU: Intensive Care Units

SIR: Susceptible-Infectious-Removed

SVM: Support Vector Machines

RF: Random Forests

LR: Logistic Regression

PMoH: Palestinian Ministry of Health

WHO: World Health Organization

HCoVs: Human COVID-19

VIF: Variance Inflation Factor

CNN: Convolutional Neural Network

MCDA: Multi-Criteria Decision Analysis

UMAP: Unified Manifold Projection

SMOTE: Synthetic Minority Oversampling Technique

LSTM: Long Short-Term Memory

PA algorithm: The Passive-Aggressive

ARIMA: Time Series Forecasting Model

AUC-ROC: The Area Under The Receiver Operating Characteristic

LDH: lymphocytes

HS-CRP: Highly Sensitive C- Reactive Protein

MLP: Multilayer Perceptron

ESR: Erythrocyte Sedimentation Rate

D-D: Differential Diagnosis Diastolic Dysfunction

ALB: Albumin Is A Protein Made By Your Liver

IL6: Interleukin 6 (IL-6), Receptor Proteins

PLR: The Platelet-To-Lymphocyte Ratio

Chapter One: Introduction and Problem Statement

1.1 Introduction

At the end of 2019, COVID-19 pandemic has appeared for the first time and started to spread widely over all countries. Since then, pandemic has returned in waves of varying frequency, either high or low, as the nature of the genetic components of the Coronavirus is constantly changing. Accordingly, the incidence of infection and the magnitude of symptoms associated with the pandemic fluctuate with each progression of waves. To date, governments have attempted to impose a range of clinical measures (e.g., construction of Intensive Care Units (ICU), quarantine) and non-pharmaceutical measures (social distancing, face mask use, large-scale closures, etc.) as well as strategies based on vaccination with different types of available vaccines, which has had a major impact on the epidemiological curve (Prem et al., 2020). However, the situation has gotten out of control in many countries due to either the lack of measures or adherence to health prevention measures and social distancing (Lei et al., 2021).

On March 11, 2020, the World Health Organization (WHO) declared COVID-19 a pandemic. Since then, governments have implemented exceptional measures in response to the pandemic, including restrictions on travel and the closure of educational institutions, businesses, and industrial facilities, resulting in substantial adverse impacts on social and economic functions. As a result, the rate of infection and the spread of the pandemic have decreased (Salman, 2020). Thus it is crucial to create sophisticated models to monitor the progression of the pandemic (i.e., infectious rate) and the severity of the symptoms.

1.2 Problem Statement

Several factors boosted the widespread of COVID-19 in Palestine such as not adhering to prevention measures or not receiving vaccinations. Most of infected people show low to moderate symptoms such as fever, cough, etc. Some COVID-19 patients needed hospitalization and ended up in ICU and ventilation machines in case of serious lung damage. In presence with

these facts the relatively weak capacity of the health care system in Palestine was unable to cope with the pandemic. Accordingly, this thesis attempts to study the phenomenon of the outbreak of COVID-19 in Palestine. Particularly, the work will use AI algorithms to model the rate of infection of COVID-19 and the severity of its symptoms.

The main concentration of this thesis is twofold. First, it aims to study the best factors that affect the likelihood of contracting COVID-19. The second aim is to study the main factors that affect the severity of symptoms for the people who are infected with COVID-19. For both goals, the thesis will study different ML models and aim to identify the best in predicting both the likelihood and severity of symptoms.

1.3 Research Questions

The main goal of this work is to study how efficiently we can predict the likelihood of infection with COVID-19 after an outbreak and the severity of symptoms in Palestine. In particular, this study aims to answer the following research questions:

Q1: What are the factors affecting the likelihood of contracting COVID-19?

Q2: What is the best model to predict the likelihood of contracting COVID-19 and why?

Q3: What are the factors affecting the severity of COVID-19 symptoms for a hospitalized patient?

Q4: What is the best model to predict the severity of COVID-19 symptoms for a hospitalized patient and why?

1.4 Importance of the study

Palestine has faced seven waves of COVID-19 so far (According to the weekly and daily report of the Palestinian Ministry of Health, 2022). Waves differed in terms of the spread rate, symptoms as well as the effect on health care, education, the economy, and other areas. Therefore, studying the behavior of COVID-19 and being able to predict the behavior of the virus depending on the clinical condition of the patient can help protect many people lives and reduce social and economic loses. This type of study contributes to providing a basis for researchers due to the lack of studies in Palestine on this topic, as this study is considered one of the first, to our knowledge, to predict the behavior of the Corona virus in the country.

Particularly, the likelihood of contracting the disease and the severity of its symptoms using AI and ML algorithms.

The advantage of using AI algorithms is its ability to provide predictions that are reasonably accurate as long as the parameters do not change. This study will represent a path and a preparation for other researchers to conduct further analysis based on the results that the study came out with. Further, this study presents a vision for the health care staff in the PMoH on how COVID-19 moves, whether in terms of its ability to spread, the strength of its spread, or its behavior. The study also provides a set of policy implications based on viable forecasts that contribute to assisting decision-makers to limit the spread of the virus, whether by strict or light government measures, while accounting for the clinical condition of patients. Furthermore, the outcomes of this study could be generalized, by decision makers, to cope with future pandemics.

1.5 Constraints of the Study

The following summarizes some of the limitations/obstacles that face this study:

- Sample boundaries: Hospitalized COVID-19 patients. No information is available about the severity of symptoms for those who contracted the pandemic and were not hospitalized.
- Temporal boundaries: The temporal boundaries of the study were limited to the first semester of the year 2021/2022.
- Spatial boundaries: Data registered with the Palestinian Ministry of Health and hospitals.

1.6 Limitations of the study

The study faced many difficulties such as the method of collecting data from the database, the limited time, the difficulty of obtaining data, and the following are the most influential determinants of the study:

1. Difficulty of obtaining data regarding the sample of patients in intensive care because the relative medical reports do not stay in the hospital system for a long time.
2. The target data was not obtained from the beginning of the pandemic due to the dispersion of data on the database on the ministry's system.

3. The sample of patients from the intensive care unit was limited in size compared to the initial sample due to the low number of patients and difficulties in obtaining their data.
4. Sample variables (testing the likelihood of Infection) were limited and there were some errors in the data notation of health personnel.

Chapter Two: Theoretical Background & Literature Review

2.1 Introduction

The recently emerged COVID-19 is one of the viruses that occur in nature on a large scale. It is a new type of coronavirus that infects the respiratory system and causes the acute respiratory syndrome. It has been called “Corona” because when examined under an electron microscope it assumes the shape of a crown, the cause of which is yet unknown (Lei et al., 2021). To this day, it is still not clear how the COVID-19 pandemic will develop due to the mutating virus. The COVID-19 pandemic not only affects the health status of individuals of a but also it affects the global public health, nation’s economy, and other social aspects. Accurately forecasting the rate of infection and the severity of symptoms posed by the virus is of paramount importance. Understanding the trajectory of the virus, such as its mode of transmission, can furnish valuable insight into how to effectively curb its spread and minimize its substantial social and economic consequences (Lei et al., 2021).

2.2 Theoretical Background

COVID-19 appeared first in late 2019 in the Chinese city of “Wuhan”. COVID-19 kept spreading quickly within the Chinese boundaries, then started to spread all over the world and It has become the most severe public health crisis since the SARS virus outbreak in China in 2003.(Aljameel et al., 2021). The COVID-19 virus is one of the corona viruses that originally infect animals. However, it can be contracted to humans and may cause harmful damage such as infecting the respiratory systems, infecting kidney cells, and many others. And when cases of infection with the virus occur from one person to another, it often occurs as a result of contact with the infected or sick person (Arti & Wilinski, 2022).

Evidence showes that this virus affects the elderly more than young and children (Glynn & Moss,2020). Moreover, the severity of COVID-19 conditions is stronger for the unhealthy as compared to healthy individuals without chronic diseases (Prem et al., 2020).

The COVID-19’s ability to spread quickly caused it to be a global threat. No country in the world was excluded from this pandemic. The pandemic has caused all groups of society to

undergo an unprecedented change in a short period, a forced change in their lifestyle that is destroying the economies of many countries, affecting health systems in all countries of the world, preventing movement and stopping flights (Song et al., 2021).

The world is in captivity to the Coronavirus, this has also reinforced what has been imposed on most of the world's population, namely strict quarantine procedures at home, travel restrictions, tests, and constant surveillance. Moreover, the spread of false information on social media, and the frightening and terrifying figures reported by various local and international media round the clock of a large number of injured and dead due to the emerging Coronavirus, leaving people in a state of panic, fear and tension on a scale that humanity has perhaps never experienced (Gozes et al., 2020). In some cases, this was also accompanied by feelings of isolation, psychological turmoil manifesting as depressive symptoms, and a general sense of boredom, which can later evolve into more severe symptoms (Gozes et al., 2020).

The COVID-19 symptoms may differ from one person to another. Some patients may not encounter any symptoms in general, some may have some low symptoms, and some other patients may have high symptoms that may lead them to hospitalization. The following are the main symptoms that affect humans when contracting COVID-19: 1. aches and discomfort, 2. Sore throat, 3. Diarrhea, 4. Eye inflammation, 5. Headache, 6. Loss of taste or odor, 7. Skin rash or discoloration of fingers or toes, to mention a few (WHO, 2019). Patients with COVID-19 recover within 14-16 days because the incubation period for the new Coronavirus is 14 days. The COVID-19 pandemic, as declared by the WHO, and its variants are highly dangerous mutant viruses that have caused a large number of deaths especially among the elderly and patients with chronic diseases (Arti & Wilinski, 2022).

After successive waves of the virus in different countries, the need for different predictive models became urgent. Countries around the world rely on such predictive models to make decisions related to the pandemic and propose new measures and evaluate the effectiveness of the measures that have been put in place. For example, Martin-Moreno (2022) argue that Long-Short Term Memory (LSTM) models may be versatile and useful, but their practical application may vary depending on the specific context and the information being analyzed. Therefore, further research is needed to compare and evaluate the performance of these models

in similar situations to determine the most reliable and practical methods for use in future outbreaks and potential pandemics.

In addition, the COVID-19 outbreaks have different trends among people. Some empirical evidence demonstrated the ability of standard models to accurately predict virus outbreaks. For example, Hsu (2020) estimated the likelihood of transmission of the virus from one person to another a model called R0, which calculates the average number of passengers on flights relative to the number of infected individuals, is used, but it is regarded as a preliminary model that does not provide a sufficient assessment of the spread of the virus. The results show that the transmissibility rate (R) has decreased, indicating that the implemented measures have effectively controlled the spread of the disease.

Moreover, important variables that must be taken into account when forecasting the spread of the virus are crucial, such as people's lack of commitment to public safety measures and social distancing rules. As a result, well-known epidemiological models, such as the curve-fitting model and the susceptible-infectious-removed (SIR) model and its extended version (eSIR) face various challenges are related to the accuracy of data, the limitations of model assumptions, the influence of human behavior, the impact of asymptomatic cases, and the effects of control measures to achieve more reliable results. One of the main challenges of these models is their ability to handle large data with high accuracy (Purkayastha & Bhattacharyya, 2021). To solve such problems, new statistical models have emerged that make different assumptions for the modeling (e.g., adding social distancing to the model, curfews, quarantines, etc.) (Samek & Müller, 2019).

Many researchers, whether in the health field or the statistical field, attempted to build appropriate mathematical models to predict the outbreak of the pandemic. When it comes to predicting the disease, there is a real problem in the availability of related information and data about the disease. As far as AI is concerned, ML, which is a type of AI that relies on historical data to make predictions, can be utilized for this purpose. But unlike other pandemic - which come and go, leaving behind useful information - there is not enough historical information about the spread of pandemic COVID-19. By definition, a pandemic is the global outbreak of a new disease. This means that, at least initially, there is no enough data to build and train a model on (Cockburn et al., 2019).

The current status of the COVID-19 pandemic continues to pose a threat to Palestine and the rest of the world. The new waves of the virus are becoming increasingly frequent and each wave presents different challenges in terms of the spread of the virus and the severity of its associated symptoms (Wu et al., 2020). Currently, there have been large numbers of COVID-19 cases detected globally, with the pandemic affecting countries all over the world, as of 2023, the COVID-19 virus, also known as SARSCoV-2, has mutated into a new strain called "BA.5." This variant, part of the Omicron lineage, is rapidly spreading across the globe and affecting countries all over the world (World Health Organization, 2023).

Sachs, et al. (2022) indicated that bolstering national health systems and increasing investments in primary and public health is crucial to cope with the pandemic. This includes investing in infrastructure, technology, and human resources to enhance health systems' ability to detect, respond to, and control outbreaks. They further recommend developing and implementing regulations for the prevention of pandemics from natural spillovers. Such measures could include early warning systems, surveillance systems, and protocols for rapid identification and response to potential pandemics. These actions would help to lower the risk of future pandemics by enabling early detection and response to outbreaks and strengthening the capacity of national health systems to effectively respond to potential pandemics (Sachs et al., 2022).

There are two main challenges in applying epidemiological models to COVID-19 in Palestine. First, the vast majority of cases go undetected because not all infected people are tested. Estimates of the percentage of undetected cases depend on the region which is evident by data collected by the PMoH. The undetected cases include unexamined primary cases that do not feel the symptoms of the disease or infected people who are not being examined, but are isolating themselves without going back to health centers (PMoH, 2021). Second, the pandemic is remarkably active over time due to various control measures. This was the main reason for the occurrence of several noticeable rises in different Palestinian areas.

The ethical aspect of using artificial intelligence algorithms in predicting the severity of COVID-19 symptoms and its spread requires careful consideration. Privacy should be a top

priority, ensuring that health data is collected and processed securely and confidentially, safeguarding it against unauthorized use or breaches. Additionally, the algorithms must be fair and unbiased in predicting symptom severity and virus spread across all social and cultural groups, avoiding any discrimination or bias. Transparency and clarity regarding how these algorithms work and the variables affecting their predictions are crucial. The processes and algorithms should be understandable and verifiable by experts and patients alike, allowing them to assess the accuracy and validity of the forecasts(Naik et al, 2022).

Furthermore, the predictions of AI algorithms should be treated as assistive tools for healthcare professionals rather than replacements for their decisions. The final medical decisions and outcomes derived from these technologies should be a collective agreement by the specialized medical team. Efforts should be made to disseminate knowledge and share the findings derived from COVID-19 symptom severity and spread predictions with the scientific community and the public. This contributes to research development and broadens the benefits in combating the pandemic.

While leveraging AI technologies in this context can improve preventive measures and pandemic responses, adhering to strict ethical principles is essential to maximize their potential benefits while avoiding any negative impact on society and individuals.

2.3 Literature Review

In this section, we will review the related previous studies that attempted to predict the likelihood of infection with COVID-19 the severity of symptoms associated with COVID-19.

2.3.1 The Likelihood of Infection with COVID-19

Various studies have looked at predicting the likelihood of infection with COVID-19 using AI algorithms. For example, the study of Marin-Gomez (2021) aimed to detect the possibility of infection with COVID-19 through a comprehensive examination using technological systems and the results of PCR tests (polymerase chain reaction), as well as analyzing the effect of concurrent factors on the probability of infection. The researcher used demographic and clinical variables recorded in the patient’s medical history and used logistic regression to indicate the probability of infection. The study was conducted on 7314 individuals

who were treated in primary care centers in Catalonia, then the decision tree was used to clarify the mechanism of the variables that affected the positive outcome of the examination. The results showed that the decision tree gave high accuracy and a good classification of the causes of injury.

Among the studies that examined the likelihood of infection and prediction of the spread of COVID-19 using logistic regression is the study by Song and Xie in 2021 aimed to forecast individuals' exposure level to the risk of infection over time using simulation methods based on the Extended Kalman Filter. The authors presented a novel approach for predicting the spread of COVID-19 that incorporates time-sensitive parameters in its estimation. Results show that the accuracy of macro-dynamic models and micro-dynamic models is limited due to the lack of detailed and comprehensive COVID-19 datasets. Additionally, while numerous models have been developed, they often prioritize short-term disease outbreak predictions and lack projections for the medium and long term. Xiong (2020) aimed to use logistic regression to predict infection and death rates of COVID-19 in California using ML. The list of independent variables includes age, gender, and ethnicity. The results showed that Latinos and African Americans had higher test-based infection rates than other ethnic groups.

Gadekallu et al. (2021) used Convolutional Neural Network (CNN) algorithm as well as the Passive-Aggressive algorithm (PA), in addition to the time series and the ARIMA model. The research was carried out on 1000 X-ray images and the results revealed that the accuracy was above 94% in Jordan, and 88% or higher in Australia. The findings indicate that deep learning, also known as AI, has the potential to be used to predict and identify COVID-19. It was also found that the disease will spread more in coastal areas, were the place of residence of the patient affects the spread of the disease. Therefore, the researchers recommended the necessity of providing aid and assistance to people who reside in coastal areas, as they are affected by humidity and high temperatures (Gadekallu et al., 2021).

Eyre et al. (2022) utilized contact testing data from England to conduct a retrospective observational cohort study on adult contacts of individuals infected with SARS-CoV-2. The researchers utilized multivariable Poisson regression to examine the relationship between transmission and the vaccination status of the infected individuals (referred to as "index

patients") and their contacts. The findings showed that 37% of the adult contacts tested positive for SARS-CoV-2 using (PCR) testing. The results also revealed that among index patients who contracted the alpha variant, receiving two doses of the vaccine was associated with a lower rate of PCR positivity among their contacts (Eyre et al., 2022).

Nordström & Nordström (2022) aimed to assess the efficacy of COVID-19 vaccination against infection, hospitalization, and death in the general population of Sweden for 9 months post-vaccination. The researchers used data from Swedish nationwide registers in a retrospective, total population cohort study. The findings suggest that there may be some differences in vaccine effectiveness between men and women, as well as between older and younger individuals. The study also discovered that vaccine effectiveness against SARS-CoV-2 infection of any severity declined over time across all subgroups, at varying rates based on vaccine type. However, vaccine effectiveness against severe COVID-19 appeared to be better sustained, although some decline was evident after 4 months. The results of this study reinforce the need for a third vaccine dose as a booster (Nordström et al., 2022).

The study conducted by Wake et al. (2020) aimed to investigate the risk of nosocomial transmission of COVID-19 in hospitals and its impact on patients with underlying medical conditions. The study took place at an NHS Trust located in South London., out of 662 hospitalized COVID-19 patients, 45 (6.8%) likely acquired the virus while in the hospital. Surprisingly, these patients did not show respiratory or influenza-like symptoms upon admission but developed symptoms and tested positive for SARS-CoV-2 through PCR testing more than 7 days after being admitted (for 38 patients, it took more than 14 days). The majority of these patients (88.9%) had shared a ward with a confirmed COVID-19 case before testing positive. To reduce the risk of nosocomial transmission, implementing a triage system that combines clinical assessment and rapid SARS-CoV-2 testing has proven effective. This system facilitates the segregation of patients, minimizing exposure to COVID-19 in shared wards. As hospitals resume regular services and potential future waves of COVID-19 admissions loom, preventing nosocomial transmission is crucial. Point-of-care diagnostic tools can aid clinical assessment by swiftly identifying COVID-19 cases, thereby decreasing transmission risk within healthcare facilities (Wake et al., 2020).

De Bruyn et al. (2022) aimed to examine the incidence, risk factors, and common pathogens associated with secondary bacterial infections in very ill COVID-19 patients. The study was conducted at the intensive care unit (ICU) of Jessa Hospital in Belgium. Among the 94 included patients, 68% acquired at least one secondary bacterial infection during their ICU stay. Secondary pneumonia was the most common infection (65.96%), followed by bacteremia of unknown origin (29.79%) and catheter-related sepsis (14.89%). Male gender, diabetes mellitus, and cumulative corticosteroid dose were identified as risk factors for secondary bacterial infections. Gram-negative bacilli were the primary pathogens in secondary pneumonia, while Gram-positive cocci were predominant in bacteremia of unknown origin and catheter-related sepsis. These findings emphasize the high incidence of secondary bacterial infections in critically ill COVID-19 patients and provide valuable insights into risk factors and pathogen profiles for effective treatment strategies (De Bruyn et al., 2022). In their study, Yang and Wang (2021) proposed a mathematical model to examine the dynamics of the transmission of COVID-19. The generated model incorporates human-to-human and environment-to-human transmission pathways. The model also incorporates different transmission rates to capture the changing epidemiological characteristics over time. The researchers focused on Hamilton County as a representative case study for the COVID-19 situation in the United States By fitting the model to publicly reported data and conducting simulations, the study revealed that the environment may play a significant role in the transmission and spread of the coronavirus. The results highlighted the importance of considering environmental factors in understanding the dynamics of COVID-19 transmission. Furthermore, the researchers used the model to simulate various epidemic scenarios and provide short-term forecasts for the development and trends of COVID-19 specifically in Hamilton County (Yang & Wang, 2021).

2.3.2 The Severity of Symptoms of COVID-19

Various studies benefit from AI algorithms to predict the severity of COVID-19. Laatif et al. (2022) used a sample of 337 patients infected with COVID-19 from Sheikh Zayed Hospital in Morocco. Both biological and non-biological data are used to predict the severity of symptoms such as blood tests, platelets, and white blood cells. Patient data was used, based on the analysis of topological data in a way called approximation and Unified Manifold Projection (UMAP). Various ML models were applied in their study with good performance encountered to help

hospitals and medical facilities prioritize patients and recommend who has a higher priority for the hospital stay based on the degree of severity (Laatifi et al., 2022).

Another study by Nemati and Ansary (2020) aimed to examine the clinical status of patients while they were hospitalized. Specifically, the study sought to forecast the survival rate and length of stay for 1182 patients. The research was based on an open-access dataset and employed seven machine-learning algorithms. The findings showed that Gradient Boosting models outperformed the other algorithms in making accurate predictions. This could be useful for healthcare providers in making decisions during the pandemic (Nemati et al., 2020).

Another study to predict and diagnose the symptoms of COVID-19 was presented by (D. Xiong et al., 2020). In their study, a dataset of 287 patients with severe and non-severe cases represented in 23 features was used. Three ML models were established using support vector machines, logistic regression, and Random Forest. Their findings indicate that Random Forest could be a useful predictive model to identify the severity of symptoms of COVID-19 (Y. Xiong et al., 2022).

Some studies examined the relationship between the severity of symptoms and survival. For instance, Aljameel et al. (2021) examined the survival rate of patients by proposing a predictive model for the early identification of COVID-19 patient outcomes through characteristics monitored while at home quarantine. A dataset of 287 patients with 20 features was used. The Synthetic Minority Oversampling Technique (SMOTE) was employed to region the class imbalance in the dataset. Three machine learning algorithms, logistic regression, random forest, and extreme gradient boosting, were utilized and compared to construct the models. The results showed an accuracy of 95%, which concludes that their model can help decision-makers and health care practitioners in the early identification of patients at risk from COVID-19 (Aljameel et al., 2021).

In the study by Chowdhury et al. (2021), machine learning was employed to develop a prognostic model for forecasting the mortality risk of COVID-19 patients. The research was conducted on 375 patients in Wuhan and utilized the XGBoost feature selection method to establish a monogram-based model. The findings indicated that the model had remarkable calibration and discrimination in predicting death probability, achieving a death probability of

80%. The authors suggested using this model for pre-disclosure and stratifying patients into low, medium, and high disease severity categories to aid physicians in making more informed decisions (Chowdhury et al., 2021).

Xiong et al. (2020) conducted a study aimed at determining a combination of four clinical indicators that predict severe or critical symptoms in COVID-19 patients. SVM algorithms were utilized and the relationship between age and protein content was examined to assess the patient's condition. The research was carried out on 336 patients. The prediction results showed a low rate of severe-critical symptoms, indicating a limited occurrence of severe or critical cases. In the study sample, patients with both severe and mild symptoms sometimes developed critical or severe symptoms (D. Xiong et al., 2020).

One of the important studies that investigated the severity of symptoms using machine learning is a study by Yan et al. (2020). The study aimed to predict criticality in patients with severe COVID-19 infection using three clinical features: a machine learning-based prognostic model with clinical data in Wuhan. A sample of 375 patients was taken where 201 patients were in good health while the rest are dead. A machine learning model was built using the Gradient Boosting (XGBoost) algorithm. The model that was built was able to forecast danger death in COVID-19 patients. In the study, the average age of patients was around 58 years, with high fever, cough, fatigue, respiratory issues, protein sensitivity, and enlarged lymph nodes as the most common severe symptoms. The utilized model displayed remarkable outcomes, revealing that lymphocytes (LDH) and Highly Sensitive C-Reactive Protein (HS-CRP) were prevalent among the cases. This model helps in forecasting the number of fatalities and the likelihood of a patient reaching a fatal stage and enables healthcare professionals to take necessary precautions such as examinations and others to prevent the spread of the disease. (Yan et al., 2020).

In Jordan, Hatmal et al (2021) predict COVID-19 symptoms using AI algorithms such as RF, XGBOOST and K-star through a clinical case study of 2213 patients who were vaccinated with various types of vaccines. It was found that the participants developed some symptoms of COVID-19, such as fever and emaciation, and the effects were strong for the elderly. RF, XGBoost, and MLP (multilayer perceptron) provided high prediction accuracies of the severity of side effects of different vaccines according to the clinical condition of patients (Hatmal et al., 2021).

The related studies were not limited to the use of AI algorithms, but parametric and non-parametric analysis methods were also used. The study conducted by Alizadeh sani et al. (2021) analyzed categorical data from 319 patients presenting symptoms of COVID-19, including fever, tightness in the chest, extreme weakness, and shivering. To compare the data between groups, a sum rank Wilcoxon test was used. The results indicated that certain medications and chronic conditions such as rheumatism, infections, and asthma were linked to the severity of fatigue and COVID-19 symptoms. However, factors such as liver and kidney disease, cough with phlegm, eczema, conjunctivitis, tobacco use, and chest pain were not found to be significant in predicting symptom severity. The study suggests that patients with chronic conditions may be at a higher risk of severe fatigue during COVID-19 infection (Alizadehsani et al., 2021).

Gui et al. (2021) collected data from 43 patients in Chongqing through its health centers. Patient demographic variables such as gender, age, and place of residence were used. Results show that age and levels of ESR (Erythrocyte Sedimentation Rate), D-D (Differential Diagnosis Diastolic Dysfunction), ALB (Albumin Is A Protein Made By Your Liver), and IL6 (Interleukin 6 (IL-6), Receptor Proteins) have a strong correlation with the status of COVID-19 patients. The ESR level is an indicator to distinguish between COVID-19 patients whose condition is considered serious. These techniques help specialists pay close attention to patients in the early stages of infection (Gui et al., 2021).

Kalem et al. (2021) used a sample of 144 patients to predict the severity of symptoms of COVID-19. They used the usual simple analysis such as the standard deviation, the arithmetic mean, and the nonparametric tests such as the Mann-Whitney test, Kruskal, the ace and chi-squared studying. The results show that the most influential variables were cough, fever, and sore throat. Based on results, the researchers recommended the necessity of therapeutic measures during the clinical treatment of patients (Kalem et al., 2021).

Mancilla-Galindo (2021) observed the retrospective effect of patients infected with COVID-19 and the relationship of age with infection severity. The study was conducted on the elderly who received health care, numbering 688 hospitals in Mexico City, and they were classified into eight groups. Using logistic regression and compared to the onset of symptoms, the elderly are more in need of medical care and more likely to die(Mancilla-Galindo et al., 2021).

Among the studies that examined the incidence of various diseases and, their relationship to the diagnosis of symptoms of infection with COVID-19 is the study of Wang et al.,(2020). They studied the level of lipids in the blood of patients with COVID-19, and then they analyzed the relationship between the level of lipids in the blood and symptoms of COVID-19. They conducted a clinical analysis of a sample of 228 patients with COVID-19 in the period of January 17, 2020, and March 14, 2020. By controlling the age and gender of the patient, results showed that the severity of symptoms with COVID-19 was associated with high levels of fats in the blood (G. Wang et al., 2020).

Jiang et al. (2020) build a framework through AI algorithms to predict the progression of the disease from mild to severe cases to provide rapid support for clinical trials. Researchers used real patient data to predict patients' acute respiratory distress syndrome (ARDS), based on data from two hospitals in Wenzhou, Zhejiang, China. The researchers concluded that there is an increase in red and white blood cells and the presence of severe pain in the muscles. Remarkably, the accuracy rate of the model was high, reaching 80% for predicting the severity of the symptoms of the disease (Jiang et al., 2020)

In a systematic review, Xiang et al (2021) show that 23 studies of the severity of symptoms of COVID-19 were based on the Bayesian method, agent-based model, and generalized growth model. This study analyzed various epidemiological parameters related to the COVID-19 pandemic. The ranges of the incubation period, serial interval, infectious period, and generation time were determined. The majority of models assumed consistency between the latent period and incubation period. Results show that travel restrictions had the most significant impact on prediction differences under different public health strategies. Contact tracking, social isolation, and improved quarantine and reporting rates were considered crucial for epidemic prevention and control. The input parameters showed significant differences in predicting the severity of the epidemic spread. Thus, caution should be exercised when formulating public health strategies based on mathematical model predictions (Xiao et al., 2021).

Xiong et al. (2022) used various ML techniques including (RF), Support Vector Machine (SVM), and Logistic Regression (LR) for predicting COVID-19 severity and to predict treatment outcomes. This model was adopted in the JinYinTan hospital and the result showed that the RF algorithm was the best in terms of accuracy to identify patients with severe COVID-19 .

Martinez & Martinez et al. (2022) used a logistic regression model with 14 variables to predict severity of symptoms in patients with COVID-19 in Mexico. The sample included 1,435,316 patients. The result showed that the model could predict the severity of COVID-19 in Mexican hospital patients.

The study by Vigon et al. (2021) conducted a meta-analysis and found that the most frequent variables used to predict severity were age, followed by immune response and vaccination. The results showed that if these variables are combined, the patient may be at a higher risk for developing kidney and liver disease, tissue swelling, and heart muscle swelling. The researchers emphasized the importance of considering factors that can lead to complications in COVID-19 patients and the need for further research on rare cases (Vigón et al., 2021)

Jain & Yuan's (2020) used meta-analysis and a sample included seven studies and focused on people with chronic diseases above 46 years. Results show that there is a clear deterioration in the clinical status of males more than females in all the studies analyzed and they were exposed to shortness of breath and even to the intensive care unit in hospitals which may end in death (Jain & Yuan, 2020).

2.3.3 COVID-19 in Palestine

Abu-Zaineh and Awawda (2022) examine the epidemiological and economic consequences of the COVID-19 pandemic. They employ a Dynamic Stochastic General Equilibrium (DSGE) model that considers variations among various population segments. The findings confirm that providing a vaccine will resolve the ongoing argument about prioritizing lives or economies. The provision of a vaccine has an immediate and positive effect both on a micro and macro scale (Abu-Zaineh & Awawda, 2021).

El-Sokkary (2021) aims to examine the most common risk factors affecting health care providers in Palestine. Clinical and epidemiological characteristics were evaluated. The silent spread of the disease among health care providers in hospitals and health centers was shown. The researchers recommended the need to re-evaluate preventive measures in health centers.

Shadeed & Alawna (2021) present a system to estimate the COVID-19 vulnerability index using Geographical Information System (GIS) and Multi-Criteria Decision Analysis (MCDA). They used 9 criteria factors (population, population density, elderly population, accommodation and food service activities, school students, chronic diseases, hospital beds, health insurance, and pharmacy). The authors created a map that highlights each governorate into COVID-19 vulnerability classes (very low, low, medium, high, very high). The developed map aims to help in making decisions in the prediction of COVID-19 in the West-Bank. Even this study helps decision-makers in highlighting which governorates are more vulnerable than others, but they did not use predictive models as well they did not use up-to-date data from the ministry of health (Shadeed & Alawna, 2021).

Chapter Three: Methodology

3.1 Introduction

In this study, AI, in particular, ML algorithms will be used to fulfill our research objectives. The following summarizes the data and models that will be used to answer each of the research questions.

3.2 Data Description

Two datasets will be used in this thesis. The first consists of all members of the Palestinian community who have taken the COVID-19 test in the West Bank during the period January 2020 to January 2022. It is worth noting that the daily number of people who take the test ranges from 2000 to 5000 persons. This dataset is obtained from the PMoH. The data provides information about individuals' demographic characteristics, the date of the test, and test results. The following summarizes the main variables included in the dataset and will be used to answer the first question.

Dependent variable: COVID-19 Status (test result). A binary variable that takes 1 if the individual is contracted with COVID-19 (true = Positive).

Independent variables:

- Age: numerical variable of the patient's age.
- Gender: categorical variable (Male, Female).
- Number of waves: (1^{st}… 5^{th}). This variable will be measured based on the date of test such that, 1^{st} from (6/3/2020 - 8/7/2020), 2^{nd} (1/8/2020 - 10/10/2021), 3^{rd} (20/10/2020 - 1/4/2021), 4^{th} (1/9/2021 - 1/12/2021), 5^{th} (25/12/2021 - 28/2/2022)
- Sample date: date of the test.
- Result date: date of tests' results.
- Cause: a categorical variable that indicates the reason for conducting the COVID-19 test (Special request/ contacts with others/having any of COVID-19 symptoms/issuance of a certificate/ traveling/ green line workers/ suspended/ medical staff/hospital admission/ transfers/ contacts/ others).
- Status: categorical variable (Follow up, New, Resampling).

- Cycle threshold (C.T.): results will be positive if the CT value is below a certain threshold (24 – 35). In general, if the value of the CT is below the threshold, then the probability of transmitting the virus will be higher.
- District: categorical variable indicating the region/governorate.
- Region (locality type): categorical variable indicating the (camp/ rural/city)
- Test type: categorical variable of the type of COVID-19 test (AG, PCR). The rapid test may give false results.

The second dataset will be used to predict factors affecting the severity of COVID-19 symptoms. This dataset is obtained from hospitals and thus contains information about hospitalized patients only. This dataset contains information about patients' characteristics such as demographics, health characteristics, and COVID-19 statistics as shown below. The following summarized variables in this dataset that will be used to answer the second question.

Dependent variable:

The two measures of severity being used in the study are:

1. Doctor-assessed Severity of COVID-19 Symptoms: This variable is present in the data and is subjective in nature, as it is based on the doctor's evaluation.
2. Researcher-calculated Severity of COVID-19 Symptoms: This is an objective measure of severity that is based on the patient's symptoms as recorded in the data and a review of related literature. This measure will take the following values: low severity; moderate severity, and high severity.

The variables that will be used to build the severity index are:

- The need for oxygen O_2: a binary variable that takes 1 if the oxygen level in the blood is less than 90%. In this case the patient must be placed on an oxygen supply device (this was used by (Aljameel et al., 2021).
- Respiration: a binary variable (high respiration $\geq$ 17 per minute for males or $\geq$ 19 per minute for females, low respiration $\leq$ 15, normal respiration = 16 for males and = 18 for females) (this was used by (Aljameel et al., 2021)..

- White Blood Cells (WBC): a binary variable (high ≥ 11, low ≤ 4.6, moderate between 4.6 – 11). A high white blood cell count indicates that the immune system is fighting off pathogens. A low white blood cell count indicates that, there is an injury or a condition that destroys cells faster than they are formed (this was used by Laatif et al. (2022)).
- Oxygen Saturation (SPO2): a binary variable in which the saturation is less than 90% (risk), and above 90% (no risk) (this was used by (Aljameel et al., 2021) & (Chowdhury et al., 2021).
- Tired: a binary variable (Yes/No) indicating if a patient is feeling tired or not(this was used by(Yan et al., 2020)).
- Shortness of Breath (SOB): a binary variable (Yes/No) (this was used by(Yan et al., 2020) & (Jain & Yuan, 2020)).
- The number of days in ICU(this was used by(Gadekallu et al., 2021).
- Temperature: categorical variable (high temperature ≥ 38.9^0 (children) and ≥ 39.4^0 (older age), (low temperature ≤ 35^0), and normal temperature (within the range $35.1^0 – 38.8^0$) (this was used by(Chowdhury et al., 2021)).
- Radiology (MRI) (ECG) or (X–ray): a binary variable (Yes/No) that shows if the patient requires an X–ray for the chest or not (this was used by(Gadekallu et al., 2021).(D. Xiong et al., 2020)
- The patient's clinical condition changes over time: categorical variable (Better/Worse/Worse than better) that measures if the condition of patients is improving or worsening over time(this was used by(D. Xiong et al., 2020)).
- Symptoms over time: categorical variable (Increased/Decreased) that shows if the patient's clinical condition has decreased or increased in severity.
- Cough: binary variable (Yes/No) (this was used by(D. Xiong et al., 2020)).
- ICU (on a bed): binary variable (Yes/No) shows if a patient needs to be admitted to the ICU (this was used by (Jain & Yuan, 2020)).

Independent variables:

- Gender: a binary variable (Male/Female).
- Age: numerical variable. Elderly people generally have more severe COVID-19 conditions.
- District: categorical variable indicating the governorate.
- High blood pressure: categorical variable in which, high blood pressure ≥ 140, low blood pressure ≤ 90, and normal blood pressure in the range 139 – 91.
- Different symptoms: Dizziness, Epigastria pain, vomiting, Loss of appetite, Fatigue … etc.
- Type of drug treatment: categorical variable (antibiotic, corticosteroid, …) showing the category of drugs a patient has received.
- Laboratory: categorical variable indicating the type of lab test a patient has undergone.
- EOSINPHILS: categorical variable (high ≥ 3, low ≤ 1, moderate between 1.1 – 2.9), which is a disease-fighting white-blood-cell and indicates if there is cancer, allergic reaction, or other types of infections such as parasitic. A high eosinophil count: a numerical variable that indicates if the body is producing, a high amount of eosinophil has to fight a bacteria, virus, or parasite. Therefore, a high eosinophil count can indicate the existence of an infection.
- Vaccinated: a binary variable (Yes/No) indicating if a patient is vaccinated against COVID-19 or not.
- Chronic diseases: a binary variable (Yes/No) indicating if a patient has a chronic disease or not.
- Type of chronic disease: categorical variable of chronic diseases (Asthma (respiratory diseases), blood pressure, diabetes, cancer, …).
- Test type: categorical variable of the type of COVID-19 test (Rapid Test, PCR, Both).

It is worth noting that the second dataset is by default a subset of the first dataset. However, the individual ID is missing from the first dataset that was given to the researcher from the PMoH. Thus merging both datasets is not a possible task.

3.3 Data exploration

After organizing the raw data, we conducted an exploration to determine the importance of variables and to verify high correlations among variables in order to prevent multicollinearity.

3.3.1 The Likelihood of Infection

The following table describes the distribution of individuals who conducted COVID-19 tests (the first dataset).

Table 3.1. The distribution of individuals who conducted COVID-19 tests

		Frequency	Percent
Gender	Male	235906	60.4
	Female	154420	39.6
	Total	390326	100.0
		Frequency	Percent
Age	Less than 25	135623	34.7
	25 to 49	176251	45.2
	50 to 75	71332	18.3
	More than 75	7120	1.8
	Total	390326	100.0
		Frequency	Percent
Result date	First wave	5147	1.3
	Second wave	82457	21.1
	Third wave	82056	21.0
	Fourth wave	95894	24.6
	Fifth wave	124772	32.0
	Total	390326	100.0
		Frequency	Percent

Status	New	385724	98.8
	Follow-up	3687	0.9
	Resampling	915	0.2
	Total	390326	100.0

		Frequency	Percent
District	Ramallah	83112	21.3
	Bethlehem	32913	8.4
	Hebron (AlKhalil)	67385	17.3
	Jenin	42972	11.0
	Jericho	18962	4.9
	Jerusalem	13895	3.6
	Nablus	63091	16.2
	Qalqilia	13469	3.5
	Tulkarm	29588	7.6
	Tubas	9927	2.5
	Salfit	14783	3.8
	Yatta	84	0.0
	Gaza	145	0.0
	Total	390326	100.0

		Frequency	Percent
Region	Urban	144041	36.9
	Rural	231830	59.4
	Camp	14455	3.7
	Total	390326	100.0

		Frequency	Percent
Test Type	PCR	338639	86.8
	AG	51687	13.2
	Total	390326	100.0

		Frequency	Percent
Result2	Injured	74113	19.0
	Not injured	316213	81.0
	Total	390326	100.0

		Frequency	Percent
Cause of test	Contact with others	174314	44.7
	Workers or travelers	84164	21.6
	Medical	22323	5.7
	Else	109525	28.1
	Total	390326	100.0

The table below shows the mean, minimum, and maximum of the numerical variables.

Table 3.2. mean values of numerical variables

Attribute	N	Minimum	Maximum	Mean	Std. Deviation
Age	390326	1.0	100.0	34.261	17.4796

Table 3.1 shows that male consists 60.4%, and female consists 39.6% of the sample, the young (–less than 49 years old) consists about 79.9% against 21.1% for old persons.

The infection rate was the lowest in the first wave, consisting of 1.3%. The second, third, fourth and fifth waves had higher percentages of positive cases, with percentages of around 21.1%, 21 %, 24.6%, 32.0% respectively. The status variable shows that 98.8% of the cases were new and 1.2% were follow-up cases or resampling.

The governorates of Ramallah, Hebron, and Nablus had the highest percentage of individual who took a COVID-19 test as compared to other governorates, with percentages of 21.3%, 17.3%, and 16.2% respectively. Individuals living in rural areas consisted about 59.4% of the sample against 36.9% and 3.7% of those living in urban areas and camps respectively.

The test type used was PCR for 86.8% of the sample and AG for 13.2%. The sample was also divided into two categories regarding injury: 19% were injured and 81% were not injured. The reasons for taking the test were categorized as: Contact with others (44.7%), Workers or travelers (21.6%), and Medical (5.7%).

3.3.2 Variables Significance Based on the Likelihood

As an initial step, visualizing the relationships between the attributes and the dependent variable helps to understand their significance. Moreover, this will help in identifying which ones may have an impact on the models. To verify that the differences in the frequency distribution are significant, we perform Chi-square tests. The results showed that all categorical variables had p-values less than 0.05, indicating that the differences are significant (in two-way frequency distribution) between each given attribute and the dependent variable.

Table 3.3. p-value for Chi-square tests variables

Variable	p-value
Gender	< 0.00
Age	< 0.00
Result date	< 0.00
Cause	< 0.00
Status	< 0.001
District	< 0.001
Region	< 0.001
Test type	< 0.001

3.3.3 Multicollinearity check

To assess the presence of multicollinearity among numerical variables, a correlation matrix was generated using Pearson's Correlation algorithm with categorical variables converted to dummy variables. The results are depicted in the following correlation matrix chart.

Table 3.4. The correlation matrix between dependent variable in likelihood of infection model (result2) and the independent variables

	Result2		
	Pearson Correlation	Sig. (2-tailed)	N/P
Gender	-.084**	0.000	Negative correlation
Age	-.028**	<0.001	Negative correlation
Result date	.074***	0.000	Positive correlation
Status	-.045**	<.001	Negative correlation
District	-.038**	<.001	Negative correlation
Region	-.006**	<.001	Negative correlation
Test type	-.110**	0.000	Negative correlation
Cause of test	0.079***	0.000	Positive correlation
***. Correlation is significant at the 0.001 level (2-tailed). N/P: Negative correlation, Positive correlation			

As indicated in the table, there is no high correlation between the categorical variables. Accordingly, none of them will be removed from the models as the Pearson correlation for all variables are between (**0.006 - 0.110**).

Table 3.5. Collinearity check using VIF test

Variable	VIF
Result2	1.016
Gender	1.009
Result date	1.027
Status	1.004
District	1.035
Region	1.052

Dependent Variable: agescale

Predictors: (Constant), Region, Result2, Status, Gender, Result date, District

To assess the presence of multicollinearity, the Variance Inflation Factor (VIF) was used after transforming categorical variables into dummy variables. A VIF value greater than 10 is generally considered an indication of high multicollinearity, while a value exceeding five may warrant concern. The VIF values in this model were below five, indicating minimal multicollinearity and no need to remove any variables.

3.3.4 Symptoms of COVID-19

The following table describes the distribution of individuals who stayed in the ICU (the second dataset).

Table 3.6. The characteristics of a sample of hospitalized COVID-19 patients

		Frequency	Percent
Hospital	Hebron Governmental Hospital	141	43.9
	Palestine Medical Complex	132	41.1
	Darwish Nazzal	12	3.7
	Jenin Hospital	6	1.9
	Beit Jala Hospital	24	7.5
	Yatta Hospital	6	1.9
	Total	321	100.0

		Frequency	Percent
District	Hebron	141	43.9
	Ramallah	132	41.1
	Qalqilya	12	3.7
	Jenin	6	1.9
	Beit Jala	24	7.5
	Yatta	6	1.9
	Total	321	100.0

		Frequency	Percent
icu_ type	Cardiac Intensive Care Department	144	44.9
	Intensive Care Department	129	40.2
	COVID Intensive Care Department	48	15.0
	Total	321	100.0

		Frequency	Percent
The need for oxygen O2	No	81	25.2
	Yes	240	74.8
	Total	321	100.0

		Frequency	Percent
Respiration	Low respiration	75	23.4
	Moderate respiration	90	28.0
	High respiration	156	48.6
	Total	321	100.0

		Frequency	Percent
White Blood Cells	Not recording	6	1.9
	Low	18	5.6

	Normal	102	31.8
	High	195	60.7
	Total	321	100.0

		Frequency	Percent
Gender	Male	198	61.7
	Female	123	38.3
	Total	321	100.0

		Frequency	Percent
Temperature	Not recording	3	0.9
	Low	9	2.8
	Normal	141	43.9
	High	168	52.3
	Total	321	100.0

		Frequency	Percent
High blood pressure	Not recording	12	3.7
	Low blood pressure	72	22.4
	Normal blood pressure	65	20.2
	High blood pressure	172	53.6
	Total	321	100.0

		Frequency	Percent
Oxygen saturation (SPO2)	Risk	189	58.9
	No risk	132	41.1
	Total	321	100.0

		Frequency	Percent
Radiology	MRI	87	27.1
	ECG	33	10.3
	ECHO	33	10.3

	XR	15	4.7
	No	153	47.7
	Total	321	100.0

		Frequency	Percent
The patient's clinical condition changes over time	Worse	117	36.4
	Worse then better	54	16.8
	Better	150	46.7
	Total	321	100.0

		Frequency	Percent
Symptoms over time	Decrease	182	56.7
	Increased	139	43.3
	Total	321	100.0

		Frequency	Percent
Tired	No	79	24.6
	Yes	242	75.4
	Total	321	100.0

		Frequency	Percent
Shortness of Breath (SOB)	No	55	17.1
	Yes	266	82.9
	Total	321	100.0

		Frequency	Percent
Cough	No	127	39.6
	Yes	194	60.4
	Total	321	100.0

		Frequency	Percent
Intensive	No	48	15.0

care unit(on bed)	Yes	273	85.0
	Total	321	100.0
		Frequency	Percent
Diagnosis	Stay ICU	162	50.5
	Leave ICU	159	49.5
	Total	321	100.0
		Frequency	Percent
Eosinophil	Low	54	16.8
	Normal	123	38.3
	High	144	44.9
	Total	321	100.0
		Frequency	Percent
Vaccine	No	198	61.7
	Yes	123	38.3
	Total	321	100.0
		Frequency	Percent
Chronic diseases	No	46	14.3
	Yes	275	85.7
	Total	321	100.0
		Frequency	Percent
Test type	Rapid	96	29.9
	PCR	84	26.2
	Both	141	43.9
	Total	321	100.0
		Frequency	Percent
Result	Died	72	22.4

	Sent back home	42	13.1
	Referred to another governmental or private hospital	48	15.0
	Improved	159	49.5
	Total	321	100.0

		Frequency	Percent
Number of days	less than 4	32	10.0
	4 to 10	104	32.4
	11 to 20	59	18.4
	More than 20	126	39.3
	Total	321	100.0

		Frequency	Percent
Type of chronic disease	No	12	3.7
	Heart disease	87	27.1
	Diabetes	28	8.7
	Liver diseases	6	1.9
	Hypothyroidism	22	6.9
	Kidney disease	11	3.4
	Lung diseases	23	7.2
	Blood diseases	72	22.4
	Orthopedic diseases	5	1.6
	Cancer	16	5.0
	Morbid obesity	6	1.9
	Not recording	33	10.3
	Total	321	100.0

		Frequency	Percent
Doctors opinion	Not very well or died	88	27.4
	Change condition or drugs	99	30.8

	Well	134	41.7
	Total	321	100.0

		Frequency	Percent
Type drugs	Antibiotic	138	43.0
	Heart drugs	48	15.0
	Else	135	42.1
	Total	321	100.0

		Frequency	Percent
Age_ MOH	Childhood (0-17)	10.00	3.12
	Adolescence (18-25)	8.00	2.49
	Youth (26-65)	177.00	55.14
	Middle-aged (66-79)	91.00	28.35
	Seniors (80-90)	29.00	9.03
	Centenarians (91 and above)	6.00	1.87
	Total	321.00	100.00

Table 3.6 describes the distribution of those infected with COVID-19 and who were hospitalized. The table shows that the majority of individuals infected with COVID-19 and hospitalized were treated at Hebron governmental hospital (43.9%) and Palestine medical complex (41.1%). The percentage of patients at other hospitals, such as Darwish Nazzal hospital, Jenin hospital, Beit Jala hospital, and Yatta hospital, is much lower, ranging from 1.9% to 7.5%. This suggests that these two hospitals may have been the primary hospitals for treating COVID-19 patients in the country.

The data in the table suggests that the district variable had a consistent percentage across all categories. For the ICU type variable, the cardiac intensive care department had the highest percentage at 44.9%, followed by the intensive care department at 40.2%, and the lowest percentage was for the COVID-19 ICU department at 15%. The table also shows that 74.8% of patients require oxygen, while 25.2% do not. In terms of respiration, 23.4% of patients require low respiration, 28% require moderate respiration, and 48.6% require high respiration. This

indicates that a majority (48.6%) of patients require high levels of respiratory support, and a significant portion (51.4%) require either low or moderate levels of respiratory support.

This data suggests that a higher percentage of individuals had high white blood cell counts. A smaller percentage of individuals had normal white blood cell counts (31.8%), and an even smaller percentage had low white blood cell counts (5.6%). The females had a lower percentage (38.3%) compared to males with the majority being male (61.7%).

This sample suggests that a higher percentage of individuals had a high temperature (52.3%) and high blood pressure (53.6%). A smaller percentage had a normal temperature (43.94%) and normal or low blood pressure (22%).

The table shows higher percentage of individuals had normal oxygen saturation (58.9%) and a lower percentage had no-risk (41.1%). In radiology, half of the patients did not do any radiology, a small percentage did ECG and echo radiology (10.3%), and an even smaller percentage did x-ray (4.7%). The patient's clinical condition change over time, about half of them showed improvement (46.7%), while a smaller percentage had worse condition (36.4%).

About half of the patients had stable symptoms over time (56.7%), while the other half had increased symptoms (43.3%) while they were in the ICU. Additionally, the results show that most of the patients were tired (75.4%), while a smaller percentage reported not being tired (24.6%).

The variable "shortness of breath" indicates that (52.9%) of patients in the sample experienced shortness of breath, while (17.1%) did not experience it. For the variable "cough," (60.4%) of the sample had a cough while (39.6%) did not. The results showed that most of the patients in the sample were still in bed in the ICU, as 85% of the sample was in that category. For the "diagnosis" variable, half of the patients stayed in the ICU, at 50.5%, while the other half, at 50%, and were discharged from the ICU.

The results of the sample indicate that a high eosinophil count was present in 44.9% of patients, while a normal eosinophil count was present in 38.3% of patients. For the "vaccine" variable, 61.7% of patients did not receive a vaccine, while 38.3% did receive one.

For the “chronic disease” variable, 85.7% of patients had a chronic disease, which is a high percentage. In terms of testing, the doctors in the hospital performed both rapid and PCR tests for 43.9% of the sample. Specifically, 29.9% of the sample had a rapid test and 26.2% had a PCR test. The results for the patients in the sample showed that 49.5% of patients improved as per the doctors’ observations, 22.4% of the patients died, and 30.2% were referred to another government or private hospital or sent home. The number of days that patients stayed in the ICU varied, starting from 3 days to 83 days. To analyze this data, patients were divided into categories. The first category, less than 4 days stayed in the ICU, represented about 10% of the sample. The second category, 4 to 10 days, represented 32.4% of the sample. The third category, more than 10 days, represented about half of the patients in the ICU.

The “age” variable for the sample patients ranges from one year to over 100 years, so it was divided into categories. The first category, less than 25 years, represented about 5.6% of the sample. The second category, 25 to 49 years, represented 18.1% of the sample. The third category, 50 to 75 years, represented 57% of the sample, and the last category, over 75 years, represented 19.3%. The age variable was categorized into six main groups based on the WHO age division, which include childhood (0-17) at 3.12%, adolescence (18-25) at 2.49%, youth (26-65) at 55.14%, middle-aged (66-79) at 28.35%, seniors (80-90) at 9.03%, and elderly (91 and over) at 1.87%. This classification transformed age from a numerical variable to a categorical variable.

The average number of days that patients stay in the intensive care unit in the hospitals under study was 19.6 days. For the “type of chronic disease” variable, the most common disease among patients was heart disease, accounting for 27.1% of the sample. The second most common was blood disease, accounting for 22.4% of the sample. Other diseases such as diabetes, high hypothyroidism, lung disease, and cancer represented a percentage between 5-7.2%.

The “doctors’ opinions” variable was divided into three categories: “not very well or died,” “change condition or drugs,” and “well.” The highest percentage, 41.7%, was for the “well” category, followed by “change condition or drugs” at 30.8%, and “not very well or died” at 27.4%. The final variable was the “type of drugs” given to patients. Antibiotics were given to 43% of the sample, and heart drugs were given to 15% of the sample.

3.3.5 Multicollinearity check

To build a model for the severity of symptoms, the correlation between the variables, which are white blood cells, the need for oxygen, and the length of stay in the intensive care unit, must be examined.

Table 3.8. The correlation matrix for the variables of Severity of symptoms

The need for oxygen O2	Pearson Correlation	.378**
	Sig. (2-tailed)	<.001
Respiration	Pearson Correlation	.480**
	Sig. (2-tailed)	<.001
White Blood Cells	Pearson Correlation	.576**
	Sig. (2-tailed)	<.001
Temperature	Pearson Correlation	.480**
	Sig. (2-tailed)	<.001
Oxygen saturation (SPO2)	Pearson Correlation	-.141
	Sig. (2-tailed)	.011
Radiology(MRI) (X-ray)(ECG)(ECO)	Pearson Correlation	.231**
	Sig. (2-tailed)	<.001
The patient's clinical condition changes over time	Pearson Correlation	-.280*
	Sig. (2-tailed)	<.001
Symptoms over time	Pearson Correlation	.279**
	Sig. (2-tailed)	<.001
Tired	Pearson Correlation	.340**
	Sig. (2-tailed)	<.001
Shortness of Breath (SOB)	Pearson Correlation	.332**
	Sig. (2-tailed)	<.001
Cough	Pearson Correlation	.470**
	Sig. (2-tailed)	<.001
Intensive care unit(on bed)	Pearson Correlation	.314**
	Sig. (2-tailed)	<.001
Number of days	Pearson Correlation	.439**
	Sig. (2-tailed)	<.001

**** The significant (p < 0.01)**

The table indicates that the Pearson correlation coefficient is good between these variables and is statistically significant.

3.4 Classification and Prediction Models

The main aim of ML algorithms is to build systems based on historical data to predict future events. ML algorithms are divided into three main categories: supervised, unsupervised, and reinforcement learning (Alloghani et al., 2020).

In this work, we will use and compare the performance of the following supervised ML algorithms: (Logistic Regression, SVM, RF, and ANNs), the following includes a brief description of these algorithms.

To build the predictive models, we will use the available datasets from the PMoH to build predictive models to predict the severity of symptoms of COVID-19. We will use the available data to split it into training/test sets. The training set will be used to train the models and the test set to test the performance of the generated models. The performance of the generated models will be measured using the Accuracy, Sensitivity, Recall, F-score and Specificity measures.

3.4.1 Binary Logistic Regression

A binary logistic regression is used in this work to model the probability of contracting COVID-19. The logistic regression depends on the process of converting the binary variable in the study into a log as follows

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where p is the probability (likelihood) of contracting COVID-19. The predictor variables $(x_1, x_2, \ldots, x_k)$ include age, gender, the number of wave, cause of conducting COVID-19 test, status, C.T., District and locality type. The sample of the corresponding datasets includes all individuals who conduct any COVID-19 test.

3.4.2 Ordinal Logistic Regression

Ordinal logistic regression is mainly appropriate when the categorical outcome of more than two classes can be ordered in a natural way such as severity status “good”, “moderate”, “bad”. This type of regression will be used to model the severity of COVID-19 symptoms. Let πj denote the ordinal probability of an observation falling in the j^{th} ordinal category (level of severity). The equation of the logistic regressions model for ordinal response is given by;

$$\log[\frac{\pi_j(x_i)}{\pi_k(x_i)}] = \alpha_{0i} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

the set of explanatory variables include (the need for oxygen O2, Respiration, White Blood Cells, Temperature, Oxygen saturation (SPO2), Radiology, the patient's clinical condition changes over time, Symptoms over time, Tired, Shortness of Breath (SOB), Cough, Intensive care unit (on bed), Number of days, doc_opinion)

3.4.3 Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are used in various ML and Data Mining applications. They can be used for classification and regression problems. SVMs work by projecting the data into higher dimensional feature spaces especially if the data is complex and/or the data is non-linearly separable (Cortes & Vapnik, 1995)

It is a method that combines statistical theory and directed education. The idea of SVMs is formulated as a search problem that divides data into two groups. The hyperplane should have the ability to separate the data whether it is linearly separable or not. If the data is not linearity separable, then SVMs will use the kernel trick to be able to divide the groups of data.

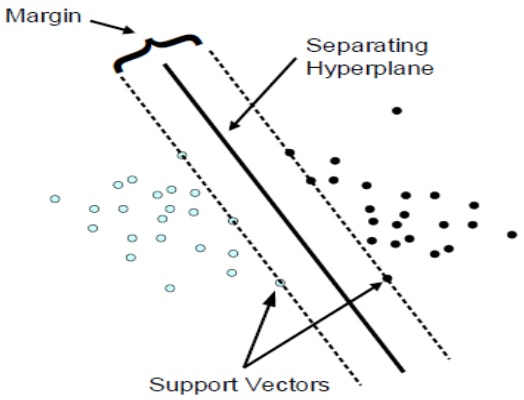

Figure 3.1. Classification using Support Vector Machines (linear separation case) (Meyer, 2015).

The task of the kernel is to project the features into higher dimensional feature spaces. However, choosing these parameters at random may lead to errors in the classification process in many cases if they are not chosen correctly (Meyer, 2015)

Various kernels exist in the literature such as the Polynomial, Sigmoid, Radial Basis Function (RBF), and many others. SVMs use training data to create separation hyperplane that separate that data into different classes. A hyperplane can be seen as a surface that separates data points falling on different sides (Bennett & Campbell, 2000).

SVMs works for binary classification problems but can be extended for multiclass classification using one-against-one and one-against-all techniques SVMs are one of the most important techniques used in data classification as they depend on multiple factors and variables that directly or indirectly affect finding the final solution. For example, the SVMs model depends on some basic parameters such as the hyperplane and Lagrange multipliers, which greatly affect the accuracy of the classification process, as the basic data in the input space are classified according to the following mathematical model:

$$w^T x_i + b \geq +1 \text{ for } d_i = +1 , i = 1, 2, \dots, N \dots\dots\dots\dots\dots\dots\dots\dots (3)$$

$$w^T x_i + b \leq -1 \text{ for } d_i = -1 , i = 1, 2, \dots, N \dots\dots\dots\dots\dots\dots\dots\dots (4)$$

where w is the weights vector, x represents the input vector, b represents the Bias value, and d represents the output value (Bennett & Campbell, 2000)

We note that the hyperplane equation is written as:

$$w^T x_i + b = 0 \dots\dots\dots\dots\dots\dots\dots\dots (5)$$

The boundary equations can also be seen in the plane as shown in figure 3.10

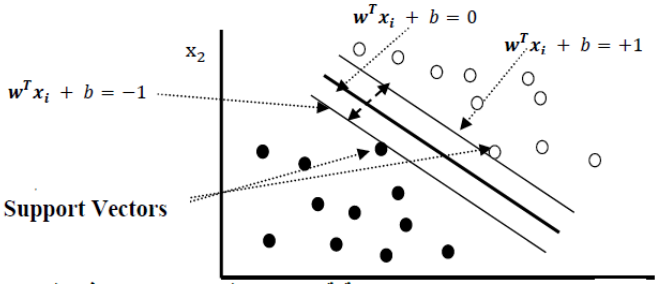

Figure 3.2. The Support Vector Machines (Ocak, 2013)

The data points that are close to or lie at the boundaries of the hyperplane are called the support vectors. The distance between points in the plane and the equation of the interval plane can also be calculated through the following relationship:

$$d(w, b, x_i) = \frac{|w^T x_i + b|}{\|w\|} \quad \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

And after mathematical transformation of the SVMs, the values of each of the ideal weights vector (w^*) and the ideal bias (b^*) are found and the classification function is calculated as follows:

$$f(x) = \text{sign}(w^*.x + b^*) \quad \ldots\ldots\ldots\ldots\ldots\ldots (7)$$

In which (w^*) represents the ideal weight, (b^*) represents the ideal bias value, and (sign) represents the final decision to belong (x) to any of the categories.

3.4.4 Random Forest (RF)

RF was introduced by (Breiman, 2001) who was inspired by earlier work by (Amit & Geman, 1997). RF is a supervised learning technique that is based on the concept of bagging, in which many decision tree classifiers are combined together to create a stronger classifier with better performance in terms of classification/regression. RFs are used for classification and regression tasks. RF has been used extensively in many applications due to its strength in applications (Pal, 2005).

A Random Forest classifier may contain many decision trees on different subsets of a given data sets. Unlike a single decision tree, the final prediction is taken from each tree and the majority of votes decide the final output.

The following figure shows how RF algorithm works.

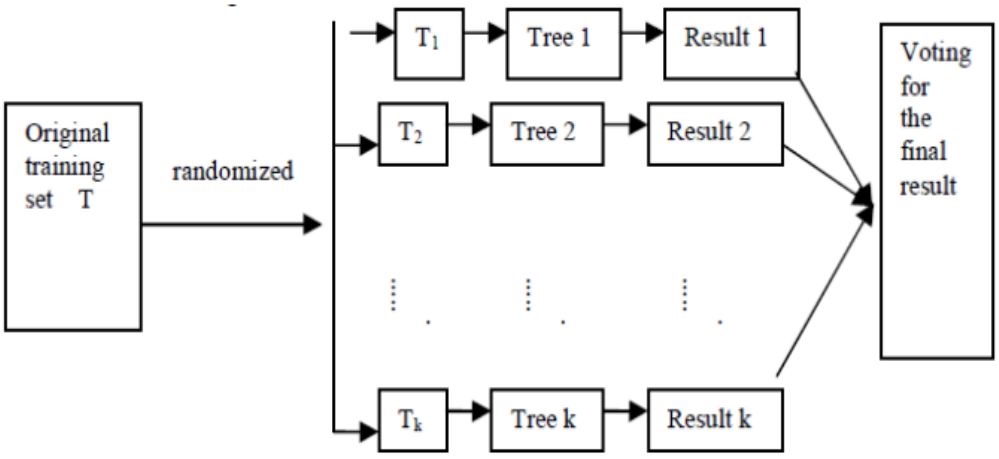

Figure 3.3. Schematic illustration of how the Random Forest algorithm for binary classification of variable (Liu et al., 2012)

"OOB" stands for "out-of-bag" and it is a method used in random forest models for estimating the accuracy of the model. In random forest models, multiple decision trees are built using a subset of the available data. Each tree is constructed using a different subset of the data, called a "bootstrap sample", which is created by randomly selecting data points with replacements from the original dataset. The out-of-bag samples refer to the data points that are not included in a particular bootstrap sample. These out-of-bag samples are used to estimate the performance of the random forest model. For each data point in the out-of-bag sample, the model is evaluated using only the trees that did not use that data point in their construction. The accuracy of the model is then calculated as the proportion of correctly classified out-of-bag samples.

Using the out-of-bag samples for model evaluation is a useful method for estimating the accuracy of a random forest model, as it provides an unbiased estimate of the model's performance without the need for a separate validation set. This can be particularly useful when the dataset is small or when it is important to use all of the available data for training the model (Bernard et al., 2012).

3.4.5 Artificial Neural Networks

Artificial Neural Networks (ANNs) are inspired by the study of the neural system (Rosenblatt, 1958).

The architecture of Neural Networks consists of several layers of artificial nodes. These layers are identified as an input layer, an output layer, and a number of hidden layer(s) that reside between the input and output layers. Each node, also called a neuron, in one layers is connected to all other neurons in the next layer. This is especially in the case of fully-connected Neural Network. This is shown in the figure 3.4 below.

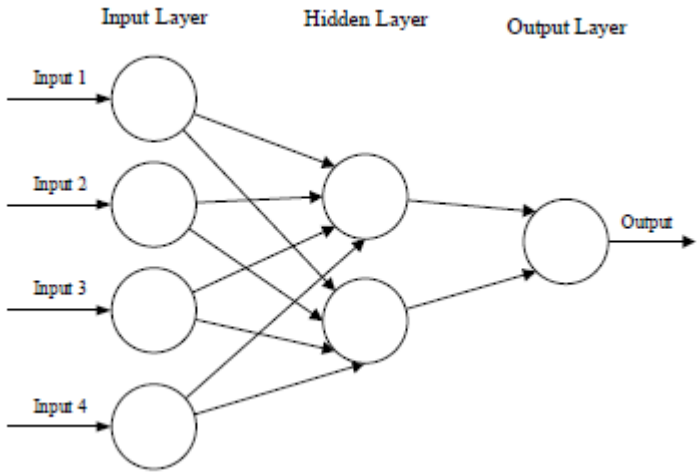

Figure 3.4. General description of a simple architecture of an Artificial Neural Network (O'Shea & Nash, 2015)**.**

Each connection between one neuron and another is characterized by its association with a value called a weight. The weights decide the importance of features and are adjusted during the learning process of the network. In its simplest implantation, a neuron multiplies each incoming input value from the neurons of the previous layer with their associated weights and then add them together. The result is then passed to a function that will decide if the value will be passed to the next layer or not. Examples of such functions include the Sigmoid, Rectified Linear Unit, and Tangent Hyperbolic functions (Ashour, 2022).

The architecture of an ANN defines the number of layers, the number of neurons in each layer, and the connections between neurons. ANN's architecture plays a crucial role in determining its performance and ability to solve a given problem. It's also important to note that different architectures may be better suited for different types of tasks or datasets. Multi-layer networks are very effective networks, especially those that include more than two layers. These networks can solve many complex problems, but they take longer to train (Wang, 2003).

3.4.6 Naive Bayes

Naive Bayes is a popular machine learning algorithm that is used for classification tasks in many fields including spam filtering, text classification, and medical diagnosis to name a few. It is based on the principle of Bayes' Theorem, which states that the probability of an event to occur is equal to the prior probability of the event multiplied by the likelihood of the event given the evidence (Raschka, 2014).

One of the key features of Naive Bayes algorithm is its simplicity and ease of implementation. The name "naive" came from the assumption that that all features are independent of each other given the target class. This is not often the case in real-world data, but the algorithm still performs well in many situations and has been widely used in practice (Gao et al., 2019).

The algorithm works by initially calculating the probability of the classes in the dataset based on their frequency in the training data. The algorithm then calculates the probability of each feature for each class using the frequency of a given feature in the training data for a particular class. Finally, these probabilities are combined to calculate the overall probability of each class given the input features. The class that results with the highest probability is then selected as the final prediction (Gao et al., 2019).

One major advantages of the Naive Bayes algorithm is its computational efficiency. It is a fast and simple algorithm that can be trained on large datasets quickly. It also requires very little data preprocessing, making it easy to use in practice(Ray, 2019).

There are a few limitations to the Naive Bayes algorithm. As mentioned earlier, the assumption of independence among features can lead to less accurate predictions in some cases. In addition, the algorithm can be sensitive to the presence of noisy or irrelevant features in the dataset.

Despite these limitations, the Naive Bayes algorithm remains a popular choice for classification tasks due to its simplicity and strong performance in many situations. It is an important tool in the machine learning toolkit and continues to be widely used in practice (Raschka, 2014).

3.4.7 Evaluation Metrics

To determine the performance of the generated models, it is necessary to use appropriate evaluation metrics to evaluate the generalization ability of the models.

There is a number of evaluation metrics including the confusion matrix, accuracy, sensitivity, specificity, precision, recall, F1-score or F-measure, and AUC-ROC curve (Jarrar, 2021).

3.4.7.1 The confusion matrix

The confusion matrix contains all information about the actual and the predicted classification.

		Predicted Class	
Actual Class		Pos	Neg
	Pos	TP True Positive	FN False Negative
	Neg	FP False Positive	TN True Negative

Figure 3.5. The confusion matrix for classification systems.

TP (True positive) is the number of correct predictions that an instance is positive. TN (True negative) represent the number of correct predictions that an instance is not positive (that is negative). FP (False positive) is the number of incorrect predictions that an instance is negative (incorrectly classified as a class of interest), FN (False negative) is the number of incorrect predictions that an instance is positive.

3.4.7.2 Accuracy

Accuracy measures the proportion of all predictions that were classifier correctly. It is used to measure the overall effectiveness of a classifier (Jarrar, 2021)

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \ldots\ldots\ldots\ldots (8)$$

3.4.7.3 Sensitivity

Also called the True Positive Rate and it measures the proportion of positive examples that were correctly classified. For example, in the health domain, the ability of the model to detect ill patients who have the conditions. It is calculated as the number of true positives (correctly classified) divided by those correctly classified (TP) and those were incorrectly classified (FN).

$$Sensitivity = \frac{TP}{TP + FN} \ldots\ldots\ldots\ldots\ldots (9)$$

3.4.7.4 Specificity

The specificity of a model is also called True Negative Rate. It measures the proportion of negative examples that were correctly classified. E.g., in the health domain, is the proportion of patients with no illness, known not to have the disease, who will test negative for it Calculated as the number of true negatives divided by the total number of negatives (TN and FP).

$$Specificity = \frac{TN}{FP + TN} \ldots\ldots\ldots\ldots\ldots (10)$$

3.4.7.5 Precision

Precision measures the proportion of true positive predictions among all positive predictions made by the model. It is calculated as the number of true positives divided by the total number of true positives plus false positives. Precision is useful for evaluating a model's performance when the goal is to minimize false positives. For instance, in a medical diagnosis where a false positive could lead to unnecessary treatments or procedures. High precision means that the model is good at not classifying as positive a sample that is in reality negative. Precision is often used along with recall to evaluate the performance of a model. Together, precision and recall can provide a more comprehensive understanding of a model's performance (Buckland & Gey, 1994)**.**

$$Precision = \frac{TP}{TP + FP} \ldots\ldots\ldots\ldots\ldots (11)$$

3.4.7.6 Recall

Recall measures the completeness of the results. It is also the true positive rate or sensitivity), It measures the proportion of positive examples that were correctly classified (from the dataset), and high recall indicates a large portion of positive examples captured in the model.

$$Recall = \frac{TP}{TP + FN} \dots\dots\dots\dots (12)$$

3.4.7.7 F-score

The F-score is a harmonic mean between the precision and recall. It has the advantage that it combines both the precision and recall in a single value.

$$F - Score = \frac{2 \times TP}{2 \times TP + FP + FN} \dots\dots\dots\dots (13)$$

3.5 Toolbox

In this study, R[1] is used as the main programming language. R offers various libraries and tools to statisticians and data scientists for loading data, data modeling, data visualization, data analysis, and ML algorithms.

The system is running on Microsoft Windows 10 Pro. It is a HP Notebook with an Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz, 2 cores, and 4 logical processors. The system operates as a x64-based PC and has 8.00 GB of installed RAM..

[1] https://www.r-project.org/

Chapter: Four Results

4.1 Introduction

There are two main objectives of this study. The first objective is to predict the probability of contracting COVID-19. The second goal is to detect the most variables that increase the severity of symptoms. The following subsections summarize the main results for both models.

4.2 The likelihood of contracting COVID-19

This section includes detailed results and accuracy measures of the classification algorithms that were used to predict the likelihood of contracting COVID-19.

4.2.1 Binary logistic regression

Binary logistic regression was applied using all independent variables. As noted in the analysis of deviance all variables were significant except the variable “Cause”.

Table 4.1. The Correlation Matrix between the dependent variable (Result2) and independent variables in the likelihood of contracting COVID-19

		Constant	Gender	Result date	C.T.	Status	District	Region	Age	Test Type
Step 1	Constant	1.000	.000	.000	-1.000-	-.811-	.000	.001	.001	.001
	Age	.001	-.030-	.028	-.001-	.000	-.006-	.021	1.000	-.052-
	Gender	.000	1.000	.000	.000	.000	-.038-	-.005-	-.030-	-.050-
	Result date	.000	.000	1.000	.000	.000	-.004-	.131	.028	-.047-
	C.T.	-1.000-	.000	.000	1.000	.808	.000	-.001-	-.001-	-.001-
	Status	-.811-	.000	.000	.808	1.000	.000	-.001-	.000	-.001-
	District	.000	-.038-	-.004-	.000	.000	1.000	-.198-	-.006-	-.031-
	Region	.001	-.005-	.131	-.001-	-.001-	-.198-	1.000	.021	.031
	Test Type	.001	-.050-	-.047-	-.001-	-.001-	-.031-	.031	-.052-	1.000

Table 4.2. Model Summary for Nagelkerke R Square between a dependent variable (Result2) and independent variables in the likelihood of contracting COVID-19

Model 1	Model 2	Model 3
20%	13%	13%

The relationship between the variables included in the model was very weak. Therefore there is no multicollinearity between them, which are variables independent of each other and do not affect each other.

Naglkerke R-squared measures the goodness of model fit. It also describes the proportion of variance that the model successfully explains. In this case it is 20%, which is considered good.

The model uses a baseline Logit Model. It represents the summary of the odds in one category relative to the baseline category which is in our case the "injured".

Confusion Matrix

The main accuracy measures are as follows:

Table 4.3. Classification table for confusion matrix for Binary Logistic Regression between the dependent variable (Result2) and independent variables in the likelihood of contracting COVID-19

Observed	Injured	Not injured	Accuracy	sensitivity	specificity
Injured	69987	4126	98.9%	99%	98.8%
Not injured	339	315874			

Binary Logistic Regression performed well in predicting the injured and since these classes are the most important in predicting the right result.

After making sure that the dependent variable is not related to the independent variables, as well as the high ability of the modulator to interpret the dependent variable (predicting the patient's injury), a simple adjustment will be made to the variables to improve the modulator's ability to predict and compare the results as in table 4.4. The table shows four Models constructed with different sets of variables. In the discussion of the table below, each model is

examined and the rationale behind variable selection is highlighted with the significance of key findings and its implications on the research.

Table 4.4. Factors affecting the likelihood of contracting COVID-19 under different set of models

Variables	Categories of Variable	Model 1	Model 2	Model 3	Model 4
		β	β	β	β
Gender	(male)	***0.312	***0.311	***0.316	**0.336
Result date	(first wave)	-22.542	-22.543		
	(second wave)	***0.192	***0.191		
	(third wave)	0.005	0.004		
	(fourth wave)	**0.751	**0.752		
Period	(wave 1 + wave2)			**-0.342	**-0.353
	(wave 3)			**-0.247	**-0.261
Status	(New)	***1.741	***1.742	***1.911	***1.901
	(Follow up)	***1.419	***1.420	***1.721	***1.711
District	(Ramallah)	0.010	0.010	*0.418	
	(Bethlehem)	-0.068	-0.065	0.351	
	(Hebron (AlKhalil))	-0.136	-0.135	0.118	
	(Jenin)	-.044	-.043	*0.442	
	(Jericho)	*0.505	*0.505	***0.793	
	(Jerusalem)	-.356	-.355	-0.073	
	(Nablus)	-.137	-.137	0.297	
	(Qalqilia)	-.146	-.144	0.286	
	(Tulkarm)	-.302	-.299	0.167	
	(Tubas)	-.082	-.081	*0.398	
	(Salfit)	-.085	-.084	*0.421	
	(Yatta)	1.452	1.452	1.521	
Governorates	Northern governorates				***0.059
	Middle governorates				***0.062
Region	(Urban)	*0.058	*0.058	***0.092	***0.099
	(Rural)	***0.154	***0.153	***0.258	***0.240
Test Type	(PCR)	**0.664	**0.663	**0.630	**0.644
Cause of test	(Contact with others)	*-0.685	*-0.685	*-0.676	*-0.689
	(Workers or travelers)	**1.088	**1.087	**1.190	**1.158

	(Medical)	*-1.071	*1.072	**1.017	*1.100
Age	Age	**-0.006	***-0.004	**0.005	**0.004
Age square	Age square		*0.000	0.00	*0.000
R^2		18.8%	20%	13%	13%
* p-value < 0.05 , ** p-value < 0.01, ***p-value < 0.001					
Period: divide the waves as: wave 1 + wave2, wave 3, wave 4+ wave 5					
The reference category is last					

Model 1 incorporates a specific set of independent variables, carefully chosen to capture relevant factors influencing the outcome. It is designed to explore the relationship between these variables and the dependent variable of interest, aiming to elucidate significant predictors. Moving on to Model 2, a different combination of independent variables is employed, reflecting an alternative approach to understanding the phenomenon under investigation. This model may focus on complementary aspects or address specific gaps in the knowledge obtained from Model 1. Model 3 takes a unique perspective by incorporating a distinct set of variables, possibly drawn from different data sources or representing additional dimensions of the studied phenomenon. This model's analysis may shed light on previously unexplored factors contributing to the observed outcomes. Finally, Model 4 offers yet another perspective, encompassing a diverse range of variables that provide a comprehensive view of the research subject. This comprehensive approach aims to integrate multiple facets and explore their collective impact on the dependent variable.

Model 1 includes all variables and shows that the variables of gender, result date, test type, cause of test, and age are statistically significant. The findings of Model 1 indicate that males have a higher probability of contracting COVID-19 compared to females. Additionally, the probability of contracting COVID-19 was higher during the second wave and fourth wave when compared to the fifth wave. The Status (New, Follow up) have a higher probability of contracting COVID-19 than the Resampling Status. Additionally, (Individuals living in Ramallah Jenin, Jericho, Tubas, Salfit) have higher probabilities of contracting COVID-19 than those living Gaza. The individuals who had a higher likelihood of COVID-19 infection were those whose diagnosis was confirmed through PCR testing, rather than those who were tested using the AG test.. Additionally, individuals living in urban areas have a higher probability of contracting COVID-19 than those living in refugee Camps. Moreover, individuals who have different reasons to test for COVID-19 have a higher probability of contracting COVID-19 compared to

individuals who only have had contact with others or for medical reasons. As well, workers or travelers have a higher probability of contracting COVID-19 compared to individuals who have different reasons to test for COVID-19. Lastly, it is reported that as age increases by one year, the probability of contracting COVID-19 decreases by 0.6%.

Model 2 includes all variables and shows that the variables of gender, result date, test type, cause of test, age, and age square are statistically significant. Model 2 differs from Model 1 only in the addition of the squared age variable to improve the model. The results are the same as in Model 1, except that R-square in Model 2 is higher, indicating that the addition of the squared age variable improves the fit of the model. It is worth noting that the beta coefficients for the independent variables in Model 1 and Model 2 are almost equivalent. This is attributed to the inclusion of the squared age variable, which helped to maintain the coherence of the beta values.

Age squared is sometimes used in binary logistic regression models as a way to capture potential non-linear relationships between age and the outcome variable. In other words, it allows for the possibility that the relationship between age and the outcome is not a simple linear one, but may have a curve or other non-linear shape.

Model 3 contains all variables with the replacement of the result date variable with the period variable, and the age variable with the squared age variable. The results were better than Model 1. Model 3 indicates that males have higher probability of contracting COVID-19 compared to females. Additionally, Based on the information provided, the probability of contracting COVID-19 is higher during the fourth and fifth waves compared to the combined first and second waves, as well as the third wave. This indicates that the later waves (fourth and fifth) have a greater likelihood of COVID-19 transmission compared to the earlier waves (first, second, and third), extending to the fifth wave.. The Status (New, Follow up) have a higher probability of contracting COVID-19 than the Resampling Status. Jericho has a higher probability of contracting COVID-19 than Gaza. The test type (PCR) has a higher probability of contracting COVID-19 than the test type (AG). Additionally, the Region (Region, Urban) have a higher probability of contracting COVID-19 than the Camp. Additionally, the individuals who have different reasons to test for COVID-19 have a higher probability of contracting COVID-19 compared to individuals who only have had contact with others or for Medical reasons. Also, workers or travelers have a higher probability of contracting COVID-19 compared to individuals

who have different reasons to test for COVID-19. Lastly, it is reported that as age increases by one year, the probability of contracting COVID-19 increases by 0.5%.

In Model 4, the governorates were categorizes into northern, central, and southern regions. Moreover, the original “district” variable was replaced with a new variable called “governorate”. The findings of this model indicated that, although the results were similar to Model 3, the probability of contracting COVID-19 was higher in the northern and central governorates compared to the other regions. Lastly, it is reported that as age increases by one year, the probability of contracting COVID-19 increases by 0.4%.

Following are the main similarities and differences among the four models:

Similarities. Model 1, Model 2, and Model 3 indicate that males have a higher probability of contracting COVID-19 compared to females. The test type PCR is consistently associated with a higher probability of contracting COVID-19 compared to the test type AG in all models. The variables “Status” (New, Follow up) consistently have a higher probability of contracting COVID-19 compared to “Resampling Status” across all models. In all models, individuals living in certain regions (Ramallah, Jenin, Jericho, Tubas, Salfit) have a higher probability of contracting COVID-19 compared to Gaza. The variables related to the reasons for testing for COVID-19 consistently show that individuals with different reasons (e.g., workers, travelers) have a higher probability of contracting COVID-19 compared to those who only had contact with others or for medical reasons.

Differences. Model 1 and Model 2 differ in the inclusion of the squared age variable in Model 2, which improves the fit of the model. Model 3 replaces the “result date” variable with the “period” variable and the age variable with the squared age variable, resulting in improved results compared to Model 1 and Model 2. Model 4 introduces the categorization of governorates into northern, central, and southern regions, replacing the "district" variable with the “governorate” variable. It shows that the probability of contracting COVID-19 is higher in the northern and central governorates compared to other regions. Additionally, the reported percentages for the increase or decrease in the probability of contracting COVID-19 with age

differ slightly across the models: 0.6% decrease in Model 1, 0.5% increase in Model 3, and 0.4% increase in Model 4.

Following modifications to the variables to improve the comprehensiveness of the study, it was observed that R^2 had converged, indicating that the standard error had also converged for all three models. Also, based on the results above (betas), the highest positive coefficient is observed for the wave 4, thus The probability of contracting covid-19 was higher during the fourth wave followed by the second. Additionally, the beta values were found to be both correct and convergent. Notably, the fourth model yielded interpretational values for the combination of age and period that improved the prediction of COVID-19, suggesting a potential relationship.

4.2.2 Naive Bayes

In this context, the Naive Bayes algorithm was used to analyze the likelihood of contracting COVID-19 based on whether individuals were infected or not. The results of this analysis were then mentioned in the statement.

The Naive Bayes algorithm, in the context of COVID-19, will predict the likelihood of an individual being infected based on their symptoms or other factors.

Table 4.5. Naive Bayes results rank the importance of predictor variables in the likelihood of contracting the COVID-19 model.

Subset	Predictor Added	Rank
1	Result	4
2	Status	5
3	District	6
4	Region	7
5	agescale	1
6	Gender	3
7	period	2
8	Cause of test	8
9	Test Type	9

The results obtained from the Naive Bayes model for the variables are consistent with the results of the logistic regression analysis, which show that the age variable is highly important and is the first variable to predict the results of the infection and the spread of COVID-19. Then

came the wave variable, which represent is the wave in which infections with COVID-19 occurred. Followed by the gender variable, then the result and status in which the test was performed, regardless if it was new or a re-test for COVID-19. The Naive Bayes model showed that the cause of the test and the type of test were the least important, ranking eighth and ninth consecutively, which contradicts the result of the regression analysis.

Table 4.6. Confusion and Accuracy Matrix for Naive Bayes of the likelihood of the contracting COVID-19

Sample	Observed	Injured	Not Injured	Accuracy	Sensitivity	Specificity
Training	Injured	56244	2506	98.2%	94.9%	98.8%
	Not Injured	3018	250492			
Testing	Injured	14076	680	98.1%	94.8%	98.7%
	Not Injured	775	62535			

The sensitivity analysis of the model revealed that it was able to accurately predict the proportion of individuals infected with COVID-19 at 94.9%. The results of the positive test were also found to be high at 98.8%, while the specificity test was able to accurately predict the proportion of individuals who were not infected with COVID-19 with a high accuracy of 98.2% in both the training and testing data. All of this indicates a high accuracy and high sensitivity of the Naive Bayes model in the prediction and classification of data according to infection and spread of COVID-19.

In a model that predicts the presence or absence of a condition or disease, high sensitivity, and specificity values indicate that the model is performing well in accurately identifying individuals who have the condition (true positives) and those who do not have the condition (true negatives), respectively.

Specifically, high sensitivity indicates that the model is able to correctly identify a high proportion of individuals who have the condition, and a low false negative rate, meaning that it rarely misses individuals who have the condition.

On the other hand, high specificity indicates that the model is able to correctly identify a high proportion of individuals who do not have the condition, and a low false positive rate, meaning that it rarely identifies individuals who do not have the condition as positive.

In general, high sensitivity and specificity values are desirable because they indicate that the model is accurate and reliable in identifying the presence or absence of the condition or disease.

4.3 The Severity of COVID-19 Symptoms

The method used to calculate the severity of symptoms of COVID-19 infection is as follows:

1) Standardize the scale for the variables that make up the severity of symptoms by measuring each variable relative to the highest value (maximum).
2) Calculate the average of the variables to ensure that the result of the severity of symptoms does not exceed one .
3) Determine the minimum and maximum value of the severity of symptoms and use the Likert triple scale to divide it into three categories: low, medium, and high severity symptoms.

4.3.1 Predicting the severity of COVID-19 symptoms

This section shows the details and accuracy measures for the classification algorithms that were used to predict the severity of symptoms.

4.3.1.1 Ordinal logistic regression

Ordinal logistic regression aims at estimating the probability of each level of the ordinal response variable (severity of symptoms) based on the values of the predictor variables.

Table 4.9. Test of Parallel Lines for ordinal logistic regression

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	3653.972			
Final	3404.616	249.356	35	<.001

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

Table 4.10. Pseudo R-Square for Ordinal Logistic Regression

Model 1 Severity (Subjective)	Model 2 Severity (Objective)
0.82%	72%

Naglkerke R-squared measures the goodness of model fit. It also describes the proportion of variance that the model successfully explains. In this case it is 20%, which is considered good.

The model uses a baseline Logit Model. This means that the model represents the summary of the odds in one category relative to the baseline category. This is in particular the case of the “High-severity Level”. The following relationship map shows how each independent variable with high correlation affects the probability of each severity level.

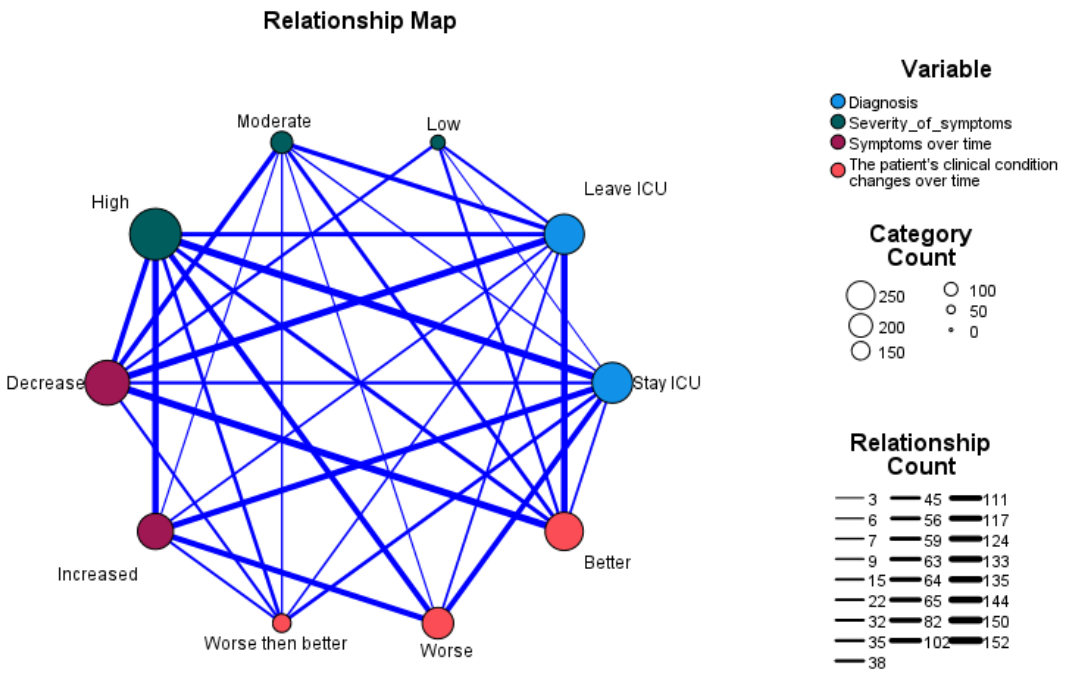

Figure 4.3. Relationship Map shows the Effect of the relationship plot of the independent variable on the severity-factor

The Effect of the relationship plot shows how the change in the independent variable might affect the response probabilities and it is based on the used model.

The severity of COVID-19 symptoms is one of the main factors that determine the need for intensive care and the need to stay in an intensive care unit.

Patients with rare and mild symptoms can be treated at home. In particular, patients with mild respiratory tract inflammation can be treated at home using the natural treatment and medication.

Patients experiencing severe respiratory tract inflammation necessitate intensive care unit treatment, while those with moderate-severity symptoms may receive treatment in a general care unit. When a patient begins to recover and their condition improves, their discharge from the intensive care unit is a manifestation of the observed progress.

The discharge from the intensive care unit depends on the breathing rate that is being dealt with and the blood pressure rate that is being dealt with. If a patient is suffering from severe COVID-19 symptoms and is prepared for treatment in intensive care unit, the treatment available in this unit can help improve the patient's condition and get rid of severe symptoms.

Intensive care treatment can be intense and requires close monitoring and management by healthcare professionals. Once the patient's condition improves and stabilizes, they may be transferred to a less intensive care setting or discharged to continue recovery at home. Follow-up care and rehabilitation may also be necessary to help the patient regain their strength and function. However, in some cases, the patient may experience severe and severe COVID-19 symptoms, in which case it may be necessary for the patient to stay in the intensive care unit for a longer period. This treatment available in the intensive care unit may require a variety of drugs and medical interventions to alleviate severe symptoms.

The severity of COVID-19 symptoms and the need for intensive care and hospitalization depend on the individual patient's condition and the specific symptoms they are experiencing.

4.3.1.2 The output of the Ordinal Logistic Regression:

Ordinal Logistic Regression performs well in predicting the high, moderate & low severity of symptoms.

Table 4.11. Confusion Matrix and Accuracy for Ordinal Logistic regression

	Predicted			
Observed	Low	Moderate	High	Accuracy
Low	84	8	2	89.4%
Moderate	11	58	9	74.4%
High	5	9	135	90.6%
Accuracy for all	31.2%	23.4%	45.5%	86.3%

The Ordinal Logistic regression model has an accuracy of 86.3%, which means that it correctly predicts the class of 86.3% of the observations in the data.

Table 4.12. Ordinal Logistics Regression is used to display the results of comparing the severity of COVID-19 symptoms predicted by the model with the severity of symptoms reported in medical reports by doctors in ICUs.

		Model 1 Severity (Subjective)	Model 2 Severity (Subjective)	Model 3 Severity (Subjective)	Model 4 Severity (Objective)	Model 5 Severity (Objective)	Model 6 Severity (Objective)
Variables		β	β	β	β	β	β
icu_ type	Cardiac Intensive Care Department	***-2.38	***-2.38	**-2.22	***-2.49	***-2.48	***-3.03
	Intensive Care Department	***-3.62	***-3.62	***-3.43	***-3.40	***-3.39	***-3.98
Gender	Male	0.29	0.29	0.41	0.12	0.12	0.15
High blood pressure	Not recording	-0.48	-0.49	-0.61	0.48	0.47	0.40
	Low blood pressure	-0.41	-0.40	-0.36	-0.77	-0.76	*-0.79
	Normal blood pressure	0.38	0.39	0.40	0.27	0.28	0.11
Eosinophil	Low	***-3.71	***-3.71	***-3.62	*-0.96	*-0.96	-0.80

	Normal	***-3.33	***-3.33	***-3.37	-0.21	-0.21	-0.07
Vaccine	No	0.15	0.17	0.29	**1.17	**1.18	**1.09
Chronic diseases	No	-2.66	***-2.66	***-2.66	-0.59	-0.59	-0.77
Type of chronic disease	No	-0.65	-0.64	-0.64	0.11	0.11	-0.01
	Heart disease	0.86	0.86	0.91	0.49	0.49	0.59
	Diabetes	0.81	0.82	0.84	0.39	0.39	0.53
	Liver diseases	-0.66	-0.65	-0.57	-0.81	-0.81	-0.62
	Hypothyroidism	0.93	0.93	0.86	0.38	0.38	0.18
	Kidney disease	0.11	0.11	0.13	0.04	0.04	-0.18
	Lung diseases	1.03	1.04	1.02	0.64	0.64	0.53
	Blood diseases	*1.10	*1.10	*0.99	0.57	0.56	0.50
	Orthopedic diseases	0.53	0.53	0.45	0.04	0.06	-0.35
	Cancer	0.67	0.67	0.54	0.52	0.52	0.56
	Morbid obesity	0.57	0.57	0.18	-0.81	-0.79	-1.66
Type of drugs	Antibiotic	-0.06	-0.06	0.01	0.43	0.43	0.52
	Heart drugs	-0.65	-0.66	*-1.00	-0.78	-0.79	**-1.04
Age		-0.01	-0.024		*0.01	0.063	
Age square			0.000			0.000	
Age- WHO	Less than 17			1.85			***23.65
	18-25			-0.51			***20.66
	26-65			0.89			***21.80

	66-79			1.38			***22.42
	80-90			0.62			21.79
Hospital	Hebron Governmental Hospital	-1.02	-1.02	-1.07	***-3.48	***-3.48	**-3.31
	Palestine Medical Complex	0.89	0.89	0.81	*-1.94	**-1.93	*-1.93
	Darwish Nazzal	20.21	20.22	19.59	***-3.51	***-3.50	**-3.90
	Jenin Hospital	-17.23	-17.21	-17.30	-21.40	-21.38	-19.35
	Beit Jala Hospital	**-4.41	**-4.41	**-4.72	***-4.04	***-4.04	**-4.61
Test type	Rapid	0.62	0.62	0.55	-0.60	-0.60	*-0.75
	PCR	***-2.13	***-2.13	***-2.11	*-0.80	**-0.80	-0.59
R^2		0.61	0.61	0.62	0.36	0.37	0.45

Table 4.12 shows the results of two sets of models. The dependent variable in the first set (Model 1 to 3) is the subjective measure of severity while the dependent variable in the second set of models (Model 4 to 6) is the objective measure of severity. The results of Model 1 (with age as continuous variable) show that there are several factors that are statistically significant in determining the severity of COVID-19 symptoms. These includes the hospital where the patient is being treated (Yatta hospital has a higher probability of severe symptoms compared to Beit Jala hospital). The department where the patient is being treated (COVID-ICU has a higher probability of severe symptoms compared to the Intensive Care Department & Cardiac Intensive Care Department). The patient's eosinophil levels high eosinophil levels have a higher probability of severe symptoms compared to low or normal levels. The type of test used (both test has a higher probability of severe symptoms compared to PCR test). And the patient has Blood diseases also have a higher probability of symptoms of severity COVID-19.

To address the potential non-linear relationship[2] between age and the severity of symptoms variable, a new variable, "Age squared," was incorporated in Model 2. The outcomes of Model 2 revealed that the beta values were similar to those obtained in Model 1. Furthermore, the statistical significance of the independent variables in Model 1 was unchanged in Model 2, and the interpretation of these variables remained the same. In Model 3, the age variable was categorized into six main groups based on the division adopted by the WHO for age. These groups include childhood (0-17), adolescence (18-25), youth (26-65), middle-aged (66-79), seniors (80-90), and centenarians (91 and above). It was discovered that there was no difference in the results compared to Models 1 and 2, except for the type of drug variation (heart drugs), which was found to be statistically significant. This showed that the severity of using other types of drugs was higher than that of heart drugs.

as for the second set of the models, the findings of Model 4 suggest that patients treated at Yatta hospital have a higher probability of experiencing severe symptoms of COVID-19 compared to those treated at Beit Jala Hospital, Hebron Governmental Hospital, Palestine Medical Complex, or Darwish Nazzal hospital. Additionally, patients in the COVID-ICU have a higher probability of experiencing severe symptoms compared to those in the Intensive Care Department or Cardiac Intensive Care Department, and individuals with high Eosinophil have a higher probability of experiencing severe symptoms compared to those with low Eosinophil. Individuals who have not been vaccinated have a higher probability of experiencing severe symptoms compared to those who have been vaccinated.

The outcomes of Model 5 revealed that the beta values were similar to those obtained in Model 4. Furthermore, the statistical significance of the independent variables in Model 5 was unchanged in Model 5, and the interpretation of these variables remained the same except Age variable was statistical significance in Model 5. Lastly, it is reported that as Age increases by one year, the probability of severity of COVID-19 increases by 0. 1%.

[2] In some ordinal logistic regression models, we have included a novel variable called "Age squared" to account for potential non-linear associations between age and the outcome variable. This additional variable accommodates the possibility that the relationship between age and the outcome may not be a straightforward linear one, but could exhibit a curved or non-linear pattern.

In Model 6, the age and age squared variables were replaced with the categorical age variable belonging to the World Health Organization. The ordinal logistic regression results showed an improvement over the results in Modules 4 and 5, except for the Eosinophil and test type variables. The statistical significance changed from PCR to rapid for the test type variable, and new variables such as the low blood pressure variable and the type of drug variable (heart drugs) were found to be statistically significant. The Age-MOH variable was particularly significant for the age groups of 0-79, as well as the test type variable (Rapid).

The findings of Model 6 suggest that patients treated at Yatta hospital have a higher probability of experiencing severe symptoms of COVID-19 compared to those treated at Beit Jala Hospital, Hebron Governmental Hospital, Palestine Medical Complex, or Darwish Nazzal hospital. Additionally, patients in the COVID-ICU have a higher probability of experiencing severe symptoms compared to those in the Intensive Care Department or Cardiac Intensive Care Department. Individuals who have not been vaccinated have a higher probability of experiencing severe symptoms compared to those who have been vaccinated.

The type of test used (both test has a higher probability of severe symptoms compared to Rapid test), the patient's Type drugs (not recording levels have a higher probability of severe symptoms compared to Heart drugs levels). Finally centenarians (91 and above) have a higher probability of severe symptoms compared to age groups of 0-79.

Differences between Objective (Models 4-6) and Subjective (Models 1-3) Measures of Severity:

Dependent Variable: The dependent variable in the first set of models (Models 1-3) is the subjective measure of severity, while the dependent variable in the second set of models (Models 4-6) is the objective measure of severity. Hospital and Department: In both sets of models, the hospital and department where the patient is being treated are found to be significant factors. However, the specific hospitals and departments mentioned may differ between the objective and subjective measures. Eosinophil Levels: Both sets of models indicate that high eosinophil levels have a higher probability of severe symptoms. However, the interpretation may differ between the objective and subjective measures. Test Type: Both sets of models suggest that the type of test used is a significant factor in determining the severity of COVID-19 symptoms. However,

the specific test types mentioned and their interpretation may differ between the objective and subjective measures. Blood Diseases and Type of Drugs: only the subjective models (Models 1-3) indicate that patients with blood diseases and using other types of drugs have a higher probability of severe symptoms. Age Variable: in the subjective models, age is included as a continuous variable (Model 1), as a squared variable (Model 2), and categorized into age groups (Model 3). In the objective models, age is included as a continuous variable (Model 4), a significant variable (Model 5), and replaced with categorical age groups (Model 6). Vaccination Status: only the objective models (Models 4-6) state that individuals who have not been vaccinated have a higher probability of experiencing severe symptoms compared to those who have been vaccinated. Additional Variables: the objective models (Models 4-6) introduce additional variables such as low blood pressure and specific types of drugs (heart drugs) as statistically significant factors in determining the severity of COVID-19 symptoms.

Similarities between Objective and Subjective Measures:

Hospital and Department: Both sets of models indicate that the hospital and department, where the patient is being treated, are significant factors in determining the severity of COVID-19 symptoms. Eosinophil Levels: Both sets of models indicate that high eosinophil levels are associated with a higher probability of severe symptoms. Test Type: Both sets of models suggest that the type of test used is a significant factor in determining the severity of COVID-19 symptoms. Age: Age is included as a variable in both sets of models, although the specific treatment of age may differ (continuous, squared, or categorical) between the objective and subjective measures. Hospital Comparison: Both sets of models compare the severity of symptoms between different hospitals, with specific hospitals having a higher probability of severe symptoms compared to others. Department Comparison: Both sets of models compare the severity of symptoms between different departments, with certain departments having a higher probability of severe symptoms compared to others. Interpretation of Significant Variables: In both sets of models, the interpretation of statistically significant variables remains consistent throughout the models, except for the additional variables introduced in the objective models. Centenarians: Both sets of models indicate that individuals in the age group of 91 and above (centenarians) have a higher probability of severe symptoms compared to the age group of 0-79.

After reviewing the selection coefficients, R-squared values, and statistical significance, it was found that Model 3 is the best model for Severity (Subjective). For Severity (Objective), it was found that Model 6 is the best model, as it provides a more accurate representation of the likelihood of increasing symptoms of COVID-19 based on additional variables, with statistical significance and a realistic function.

4.3.1.3 Support vector machines

SVMs algorithm maps data points into a higher dimensional feature space and then finds the hyperplane that maximizes the margin between data points. The radial basis function (RBF) kernel is often used to map data especially if the data is not linearly separable. In this work, we use a 10-fold cross validation, which is a common method for evaluating the performance of a model as it allows for a more robust estimate of model performance by training and testing the model multiple times with different subsets of the data. Both parameters of the RBF kernel (the Cost and epsilon) are estimated through the cross-validation process. The cost parameter C controls the tradeoff between increasing classification accuracy and simplifying the complexity of the model and the epsilon parameter determines the width of the of the kernel function.

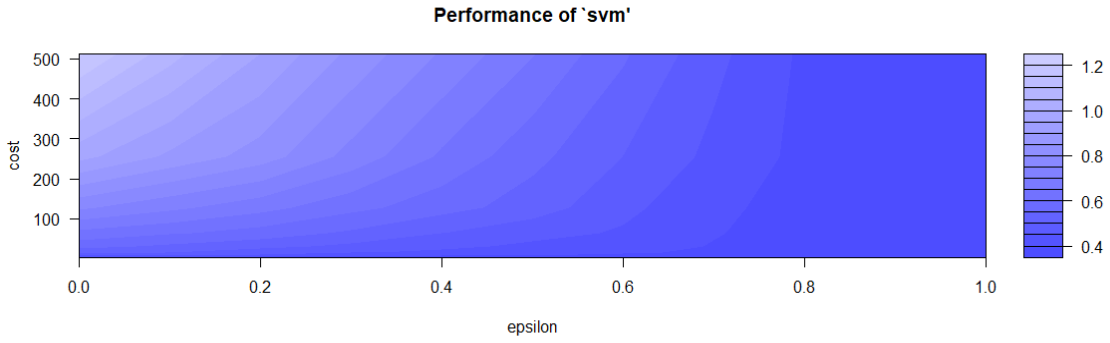

Figure 4.4. The cost parameter C controls the tradeoff between maximizing classification accuracy and minimizing the complexity of the model

It appears that as the value of the C parameter increases in the training data, the value of the error decreases, indicating that the SVM model is a good model for data with severe symptoms of COVID-19. This is because a high C value puts a stronger emphasis on correctly classifying the training data, which can lead to a model that generalizes well to unseen data with similar characteristics.

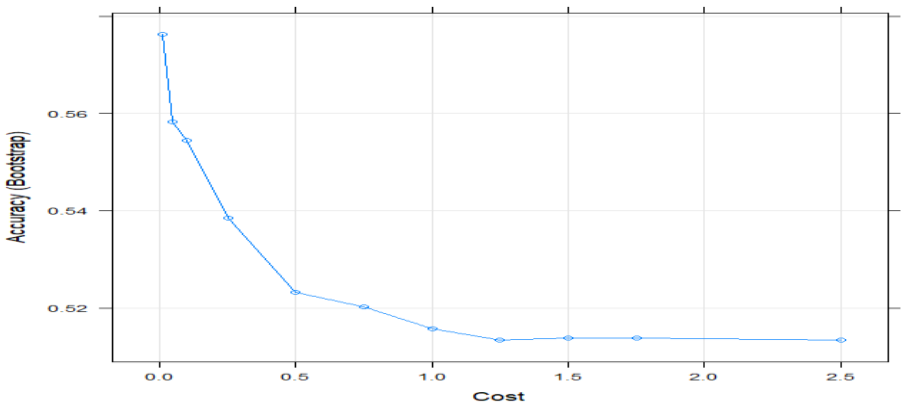

Figure 4.5. The effect of changing the cost parameters on the accuracy of the model using the RBF kernel

The confusion matrix and the accuracy measures were as follows

Table 4.13. Confusion Matrix the Accuracy measures for SVMs

Measures	Low	Moderate	High
Sensitivity	0.0	70%	93%
Specificity	1.0	82%	60%
Balanced Accuracy	50%	70%	81%

The model can accurately predict the outcomes for most of these patients based on the variables included in the model, which predicts with very good accuracy through SVMs the severity of the patient's symptoms.

4.3.1.4 Neural Network Classifier

The analysis is based on the available data with few initial relationships, using neural network that shows relationships between variables and models based on the weight of each variable. Using this model to predict the variables that most affect the severity of COVID-19 symptoms

Table 4.7. Independent Variable Importance to predict the variables that most affect the severity of COVID-19 symptoms

Independent Variable Importance	Importance	Normalized Importance
The need for oxygen O2	.066	34.7%
Respiration	.084	43.8%
White Blood Cells	.073	37.9%
Temperature	.122	63.8%
Oxygen saturation (SPO2)	.056	29.3%
Radiology(MRI) (X-ray)(ECG)(ECO)	.094	49.2%
The patient's clinical condition changes over time	.052	27.3%
Symptoms over time	.039	20.5%
Tired	.057	29.9%
Shortness of Breath (SOB)	.049	25.5%
Cough	.043	22.2%
Intensive care unit(on bed)	.073	38.3%
Number of days	.191	100.0%

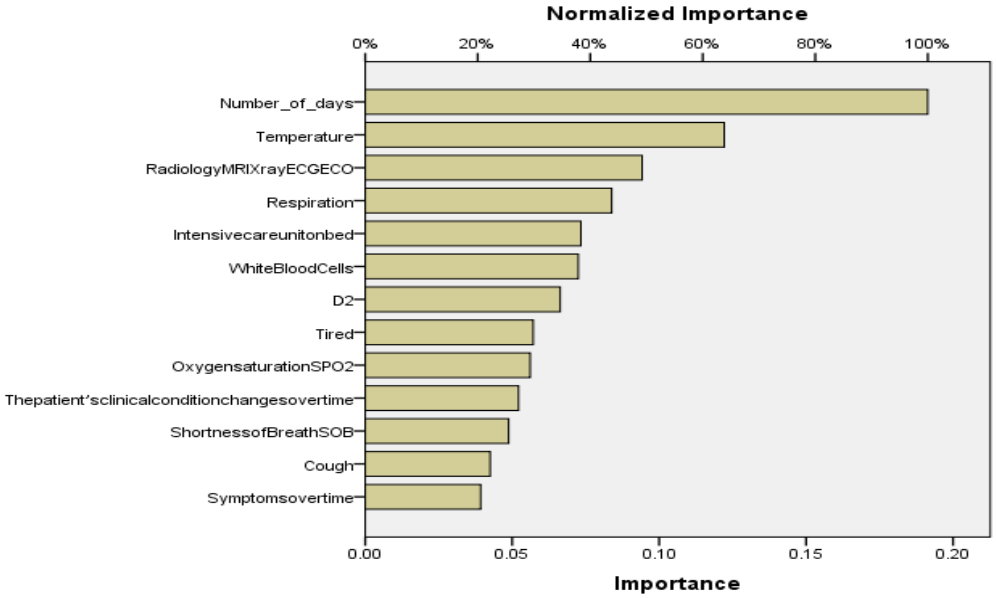

Figure 4.1. Independent Variable Importance to predict the variables that most affect the severity of COVID-19 symptoms

The neural network has determined that the number of days, temperature, and various radiology results (MRI, X-ray, ECG, ECO) are the most important variables in predicting the severity of COVID-19 symptoms, while respiration and symptoms over time are also important. The least important variable appears to be SPO2, which may not always be a reliable indicator of symptom severity. This suggests that using all of the variables in the model is important for

accurate predictions. Additionally, it highlights the importance of considering the number of days of stay in the ICU as a key factor in determining the severity of COVID-19 symptoms.

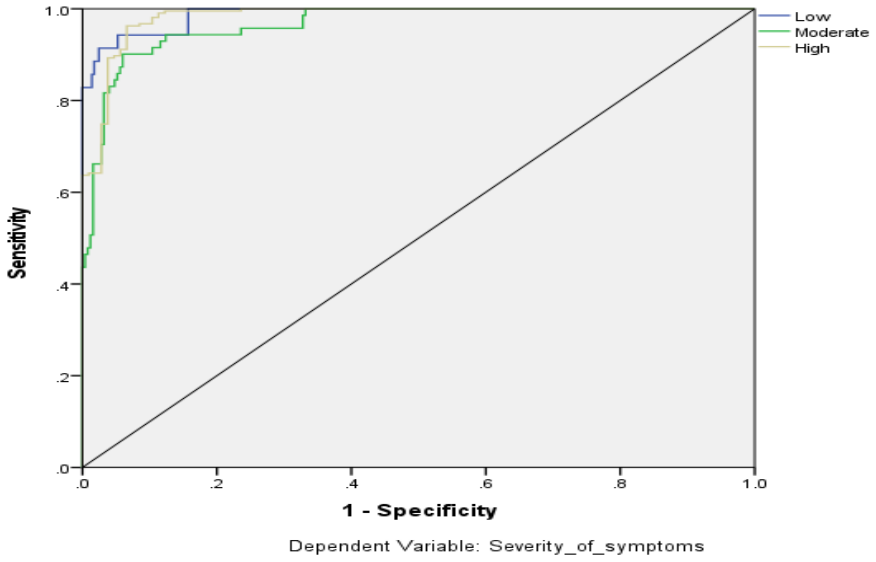

Figure 4.2. Sensitivity and Specificity for ANN Classifier in the severity of COVID-19 symptoms

The high prediction rate of 89% from the sensitivity and specificity tests suggest that the ANN algorithm is an effective method for predicting the severity of COVID-19 symptoms using the identified variables. This high prediction rate indicates the potential usefulness of this model for classifying patients into mild, moderate, and severe symptom categories, and for predicting the severity of symptoms in new cases. The ability of the model to classify data into low, moderate, and high severity based on the independent variables highlights the importance of using this model to aid in patient care and treatment decisions.

Table 4.8. Confusion and Accuracy Matrix for Neural Network of the severity of COVID-19 symptoms

		Classification			
Sample	Observed	Low	Moderate	High	Accuracy
Training	Low	15	3	0	83.3%
	Moderate	0	42	7	85.7%
	High	0	3	143	97.9%
	Accuracy	7.0%	22.5%	70.4%	93.9%

Testing	Low	13	4	0	76.5%
	Moderate	0	18	4	81.8%
	High	0	2	67	97.1%
	Accuracy	12.0%	22.2%	65.7%	90.7%

Dependent Variable: Severity_of_symptoms

Based on the results of the model, it appears that ordinal logistic regression can be used to predict the severity of COVID-19 symptoms. This is consistent with previous studies that have identified the importance of white blood cell count, respiratory function, and oxygen needs as indicators of symptom severity. The high prediction rate from the model, along with its ability to classify symptoms into mild, moderate, and severe categories, suggests that ordinal logistic regression can be an effective tool for predicting COVID-19 symptom severity. Additionally, it is important to note that other variables such as days of stay in ICU, temperature and radiology results, symptoms over time and SPO2 also play an important role in determining the severity of symptoms.

4.3.1.5 Random Forest

In Random Forests, a collection of decision trees is created to infer the most important variables that entered to the model. The following models are tuned in RFs:

mtry: the number of variables that are randomly sampled as candidates at each split.
ntree: the number of trees to grow in the model.

These parameters have the highest effect on model performance and the following figure shows the error for all dependent variable classes:

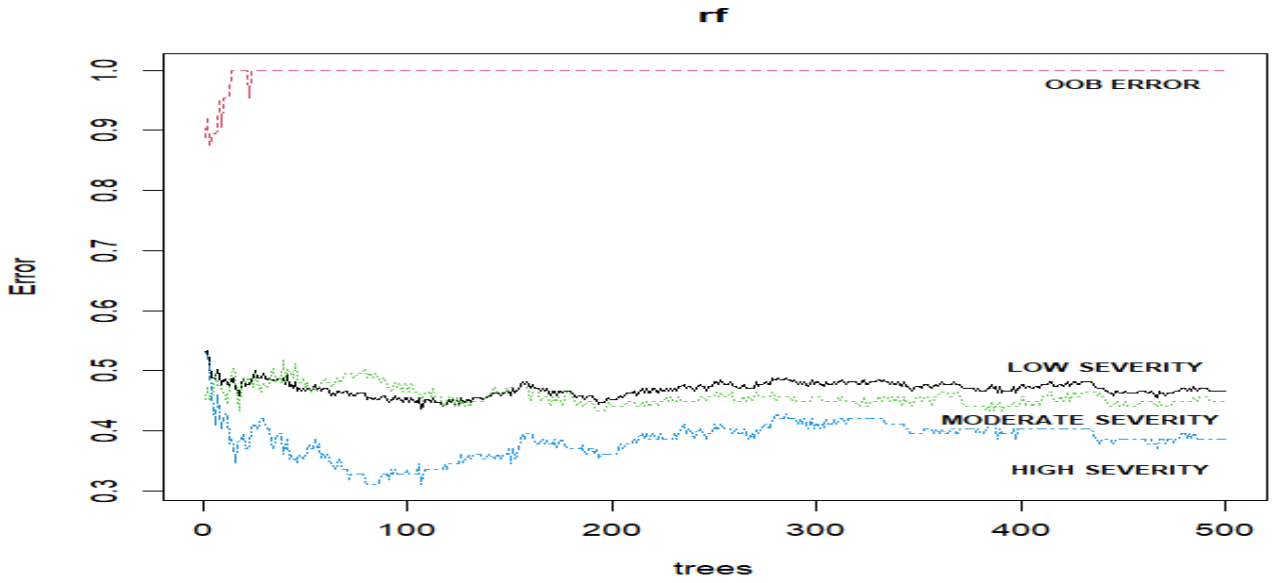

Figure 4.6. Error based on the number of trees in Random Forest stabilizes after a few trees, so choosing 400 or 500 trees is good enough.

There is a relation between the number of trees and the error in the Random Forest. When the number of trees in the Random Forest is low, the error resulting from the number of trees in the Random Forest is relatively higher in terms of accuracy. On the other hand, if the number of trees in the Random Forest is high, the error resulting from the number of trees in the Random Forest is relatively lower in terms of accuracy.

As shown in the previous figure, the error resulting from the number of trees decreases as the severity of COVID-19 symptoms increases. That is, the number of trees in the Random Forest, which is formed due to the large number of variables used in creating the tree in the case where the severity of COVID-19 symptoms is high. Accordingly, the error increases as the severity of COVID-19 symptoms decreases. That is, the Random Forest algorithm was able to predict the lowest possible error with the strength of COVID-19 symptom severity. The best value of the parameter mtry to set is 2 since it has the least OOB Error = 0.091.

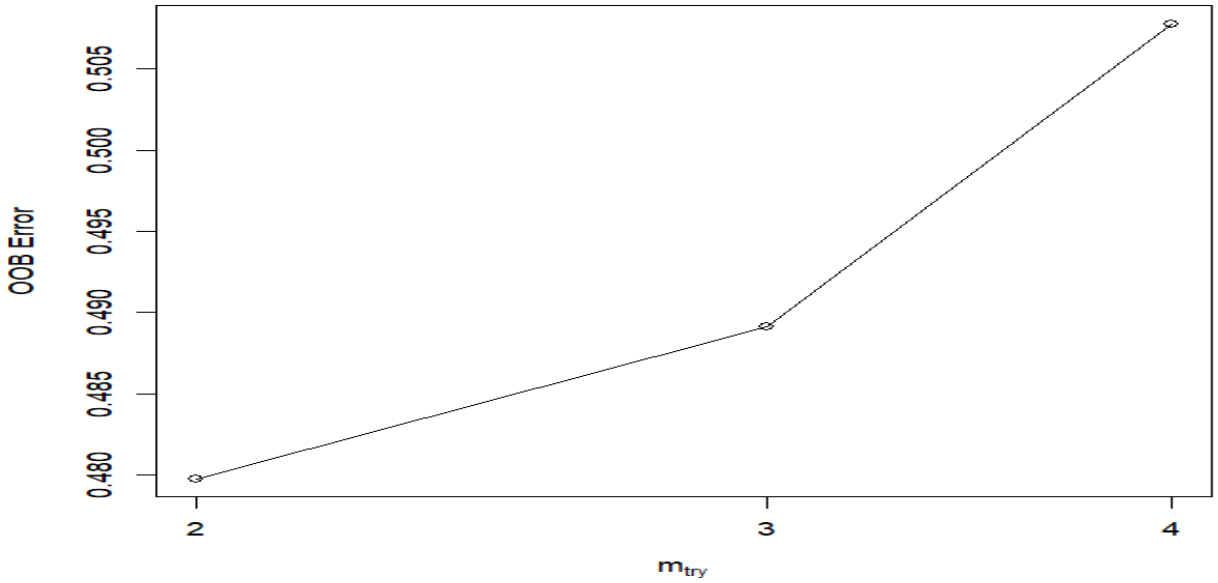

Figure 4.7. Error based on mtry in Random Forest

It appears that the lowest OOB error used by the Random Forest algorithm was 0.091, indicating a high prediction accuracy for the severity of COVID-19 symptoms using the current variables. The closer the OOB error is to zero, the higher the prediction accuracy. This suggests that the independent variables were very useful in predicting the severity of COVID-19 symptoms in intensive care, based on the available data.

The confusion and the accuracy measures are as follows:

Table 4.14. Confusion Matrix & The Main accuracy measures for Random Forest

Measures	Low	Moderate	High
Sensitivity	86%	86%	96%
Specificity	99%	94%	87%
Balanced Accuracy	84%	90%	92%

Figure 4.8 shows the mean decrease Gini-index which measures how much the model fit decreases when a variable is dropped. The greater the drop the more significant the variable is. In other words, it shows the importance of each variable in the generated forest.

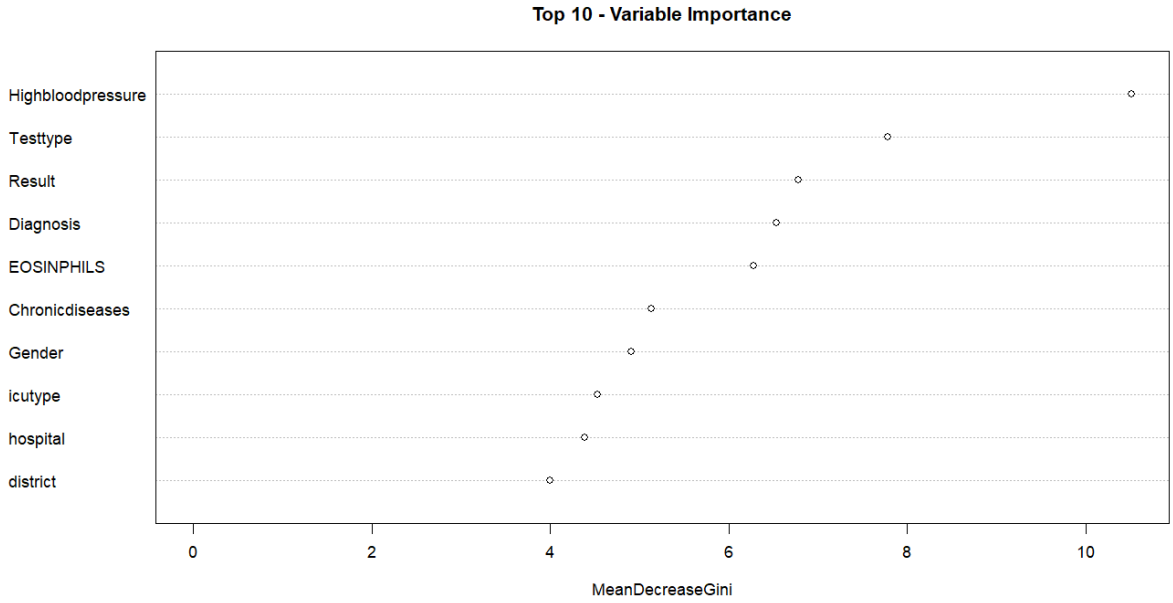

Figure 4.8. Mean Decrease Gini in RF for variable importance

The results of the Random Forest algorithm show that after performing the appropriate classifications and divisions for decision trees, the accuracy is high and it was able to predict the most influential variables in the model.

These variable should be taken seriously into consideration as they help doctors diagnose the patient's condition, give the patient the appropriate medication. That was conducted based on Mean Decrease Gini in Random Forest for variable importance.

We found that the most important variables for doctors to take seriously are the patient's condition if they stay in the ICU, and then the test result and the patient's condition reaching a specific development that can lead to death or hospitalization. The importance of diagnosing severity of symptom and the high blood pressure then test type, and the presence of other diseases in COVID-19 patients.

The results suggest that both ordinal logistic regression and Random Forest algorithm were able to predict symptom severity, with the latter being more accurate. The variables used by doctors were found to be important by both models, and contributed to the prediction of symptom severity.

We use Random Forest for predicting the severity of COVID-19 symptoms because the algorithm relies on the Gini index for data splitting and feature selection. The Gini index evaluates the impurity in the dataset and minimizes the randomness, allowing the model to make more accurate predictions. Additionally, Random Forest utilizes the out-of-bag (OOB) samples during training, which acts as a validation set, providing an unbiased estimate of the model's performance without the need for a separate test set. This enables the model to generalize well to unseen data and enhances its overall predictive capabilities for COVID-19 symptom severity.

Following is a table those summaries the accuracy of all classifications:

Table 4.15. Confusion Matrix & the accuracy measures for the likelihood of contracting COVID-19

Models	Accuracy	Sensitivity	Specificity
Binary Logistic Regression	98.2%	99.6%	99.4%
Naive Bayes	98.9%	94.9%	98.8%

The table shows that comparison of Accuracy, Sensitivity and Specificity across COVID-19 Contracting Likelihood Models: Naive Bayes Model Emerges as Best Performer with Variable Importance Ranking.

Table 4.16. Confusion Matrix & the accuracy measures for the severity of symptoms of COVID-19

Models	Accuracy	Sensitivity	Specificity
Artificial Neural Network	90.9%	89%	89.1%
Ordinal Logistic Regression	86.3%	84.1%	80.2%
Support Vector Machine	81%	81.1%	80.2%
Random Forest	91.7%	95.8%	85.2%

The table shows that assessment of COVID-19 Symptom Severity Prediction Models: Random Forest Model excels in Accuracy, Sensitivity, and Specificity with Variable Importance Ranking.

Chapter Five: Discussion and Conclusion

The previous chapter presented the analysis of data and the details of using Machine Learning algorithms in predicting the likelihood in contracting COVID-19 and to measure the severity of symptoms. The results indicated good performance in-terms of performance measures. This chapter will further interpret the results and policy implications will be made based on the findings.

5.1 Discussion

In this section, the results of each part of the study will be discussed in details. The discussion will begin with the results of the part related to the probability of contracting COVID-19.

Based on the results of the sample of individuals for the likelihood of infection with COVID-19, it was found that males have a higher probability of contracting COVID-19 than females. However, the exact reasons for this are still being studied and are not yet fully understood. Bwire (2020) stated that it could be due to various reasons such as sex hormones and different life style to men compared to women. Moreover, behavioral differences could be another factor. Males and females may differ in their behaviors, which could affect their likelihood of contracting the virus. For example, males may be more likely to engage in risky behaviors, such as not wearing a mask or social distancing, that could increase their exposure to the virus (Sanz-Muñoz et al., 2021). Lastly, occupational differences: males and females may also differ in the types of jobs they have and some occupations may put individuals at higher risk of exposure to the virus. For example, males are more likely to work in healthcare settings or jobs that require frequent interactions with others, which could increase their risk of contracting the virus (Coombs, 2020). However, this is still a debatable issue the main causes of why men have higher chance of contracting the disease the women is out of the scope of this research.

Another result emerging from this study is worth highlighting is related to the type pf test. Individuals who conducted the PCR tests have a higher likelihood of detecting the virus than those who conducted the AG tests. This explains the reason may be that PCR (Polymerase Chain Reaction) tests are considered to be more accurate than AG (antigen) tests in detecting the virus

because they detect the genetic material of the virus, while AG tests detect specific proteins on the surface of the virus. PCR tests are able to detect very low levels of the virus, even if the person being tested is asymptomatic, while AG tests are more likely to produce false negative results. Additionally, PCR tests are able to detect if the virus is present in the early stages of the infection, while AG tests may not detect the virus until the person is already showing symptoms (Viloria Winnett et al., 2022).

Additionally, individuals with specific reasons for testing, such as being workers or travelers, have higher likelihood of contracting the virus compared to those who have had contact with others because they are more likely to come into contact with a larger number of people in different settings and places, which increases the risk of being exposed to the virus. For example, workers in certain industries such as healthcare, transportation, and essential services are more likely to be exposed to the virus as they are in contact with many people, while travelers are more likely to come into contact with the virus during their travels. Also, individuals who travel to high-risk areas or have contact with high-risk groups, have a higher likelihood of contracting the virus compared to those who have had contact with others.

The probability of contracting COVID-19 increases as age increases because as people get older, their immune systems become weaker and less able to fight off infections. Additionally, older adults are more likely to have underlying health conditions such as heart disease, diabetes, and lung disease which can make them more susceptible to severe illness if they contract COVID-19. Furthermore, older adults are more likely to live in congregate settings such as nursing homes, where the virus can easily spread among residents (Alexander et al., 2021).

Based on the results above, the probability was higher in the fourth wave for several reasons. The first reason is that the virus has mutated and new variants have emerged which may be more transmissible and more resistant to current vaccines (L. Wang & Cheng, 2022). These new variants spread more easily and quickly than previous strains, leading to more people getting infected. Another reason is that many people have become complacent and less vigilant in following guidelines for preventing the spread of the virus, such as social distancing and mask wearing. This can lead to more interactions between people and more opportunities for the virus to spread. Also, some countries have eased the restriction and opened up their economies, leading to more gatherings and social interactions. Finally, the increasing availability and

distribution of vaccines, while helping to protect many people, has also led to a false sense of security among some, who may be more likely to engage in high-risk behaviors.

Turning to the results related to the severity of symptoms, many points are worth emphasizing. Patients treated at Yatta hospital have a higher probability of experiencing severe symptoms of COVID-19 compared to those treated at Beit Jala Hospital, Hebron Governmental Hospital, Palestine Medical Complex, and Darwish Nazzal hospital. There could be various factors that contribute to this difference such as the population that the hospital serves, the level of care provided, the availability of resources, and the variation in patient demographics and underlying health conditions. It is important to conduct a thorough investigation and analysis to determine the causes for this difference in order to provide targeted interventions and improve outcomes for patients.

Additionally, patients in the COVID-ICU have a higher probability of experiencing severe symptoms compared to those in the Intensive Care Department or Cardiac Intensive Care Department because they are specifically being treated for COVID-19, which is a highly infectious disease that can cause severe illness, including respiratory failure. The patients in the COVID-ICU have more advanced cases of illness and are more likely to require intensive care and specialized equipment. Additionally, the COVID-ICU is designed to take precautions to minimize the risk of virus spreading within the hospital, such as separating COVID-19 patients from other patients and implementing strict infection control measures, which can also contribute to patients experiencing more severe symptoms. Furthermore, patients in the COVID-ICU are more likely to have underlying health conditions that put them at a higher risk of severe illness. Furthermore, patients who have been diagnosed to leave ICU have a higher probability of experiencing severe symptoms compared to those who are staying in ICU, because the ICU is a high-risk environment that requires close monitoring and management of patients to ensure they are receiving optimum care during their stay. However, some patients are discharged from the unit prematurely, even though they may have a prolonged stay in the hospital.

Another important result is related to the vaccination. Individuals who have not been vaccinated have a higher probability of experiencing severe symptoms compared to those who have been vaccinated, because vaccines work by training the immune system to recognize and fight the virus. When a person is exposed to the virus, their immune system can quickly produce

the necessary antibodies to prevent or lessen the severity of the disease. This is why vaccines are one of the best ways to protect patients from severe illness caused by COVID-19. Additionally, vaccines are particularly effective at protecting vulnerable individuals, such as the elderly and those with underlying health conditions, from severe illness and death.

Individuals with high blood pressure (i.e., hypertension) may have a higher probability of experiencing severe symptoms from COVID-19 because hypertension can increase the risk of complications from viral infections such as pneumonia and Acute Respiratory Distress Syndrome (ARDS). High blood pressure can also put added stress on the heart and blood vessels, making it harder for the body to fight off the virus. Additionally, hypertension is a risk factor for other underlying health conditions such as diabetes and heart disease, which can also increase the risk of severe symptoms from COVID-19.

Individuals with high Eosinophil have a higher probability of experiencing severe symptoms compared to those with low Eosinophil, because high levels of eosinophils have been associated with severe COVID-19 symptoms and it is thought that eosinophils play a role in the inflammatory response to the virus. A heightened inflammatory response may contribute to the severity of symptoms. Additionally, eosinophils have been shown to have a role in viral infections and the immune response to them, and it is possible that individuals with high levels of eosinophils may have a stronger immune response to the virus that leads to more severe symptoms.

If a patient does not have any recorded medical history by the doctors or nurses it could make it more difficult for the healthcare team to identify underlying health conditions that may put the patient at a higher risk for severe symptoms of COVID-19. Having a history of certain conditions such as heart disease, diabetes, hypothyroidism, lung diseases, or cancer can indicate that a patient may be at a higher risk for complications if they contract COVID-19. These conditions can affect the body’s ability to fight off the virus, or they may be associated with a higher risk of developing severe symptoms. Additionally, having a recorded medical history allows the healthcare team to more quickly and effectively make treatment decisions that could potentially reduce the risk of severe symptoms. Furthermore, patients who are taking other drugs prescribed by the doctors have a higher probability of experiencing severe symptoms compared to those who are taking heart drugs. It is possible that the patients who are taking other drugs

prescribed by the doctors have underlying health conditions, which are associated with a higher risk of severe symptoms of COVID-19. It is important to consider all the medications that a patient is taking when making treatment decisions, as some drugs may interact with others and have an impact on the patient's overall health.

Regarding age, there are several reasons why the elderly may have a higher severity of symptoms of COVID-19 infection. Firstly, as people age, their immune system becomes weaker, making it harder for them to fight off infections. Additionally, elderly individuals are more likely to have underlying health conditions that can make them more susceptible to severe symptoms of COVID-19.

5.2 Conclusion

This thesis highlights the importance of using machine learning in the health sector, particularly in the context of COVID-19. The study suggests that using machine learning algorithms can help predict the level of spread of the disease in Palestine and identify the most important variables that can help predict the severity of COVID-19 symptoms in the ICU unit in Palestinian hospitals. By automating the process, hospitals can receive real-time alerts when extreme values occur that indicate high symptom severity, allowing them to take immediate action.

This approach can help save time and increase the accuracy of detecting symptom severity, as the speed of the spread of COVID-19 is based on a set of predicted variables. The study also showed that most of the models accurately predicted the strongest predictive variables, with accuracy values ranging between 80% - 99%, which indicate the strength and accuracy of the models.

This study has provided valuable insights into the factors associated with the likelihood of contracting COVID-19 and the severity of the infection. The findings suggest that males, older individuals, and those with specific reasons for testing are at a higher risk of contracting the virus. The study has also highlighted the importance of using PCR tests for accurate detection of the virus and the need for targeted interventions to improve outcomes for patients in hospitals

and hospital departments. As the pandemic continues to evolve, further research and analysis will be necessary to better understand the virus and develop effective strategies to mitigate its impact. Moreover, this study highlights the importance of data and tracking data of patients in the healthcare sector as it showed how data collected during the pandemic can help in predicting various aspects of the outbreak (in our case the likelihood of contracting the disease and the severity of symptoms), which would help in decision makers to take the correct course of action and for better planning. Furthermore, this study can be used as a guideline for future studies in case of similar pandemics by utilizing new available data and the predictive power of machine learning algorithms to possibly predict the behavior of new outbreaks.

References

Abu-Zaineh, M., & Awawda, S. (2021). *Assessing the Health and Economic Impact of the COVID-19 Pandemic in Palestine*.

Alexander, P. E., Armstrong, R., Fareed, G., Lotus, J., Oskoui, R., Prodromos, C., Risch, H. A., Tenenbaum, H. C., Wax, C. M., & Dara, P. (2021). Early multidrug treatment of SARS-CoV-2 infection (COVID-19) and reduced mortality among nursing home (or outpatient/ambulatory) residents. *Medical Hypotheses*, *153*, 110622.

Alizadehsani, R., Alizadeh Sani, Z., Behjati, M., Roshanzamir, Z., Hussain, S., Abedini, N., Hasanzadeh, F., Khosravi, A., Shoeibi, A., Roshanzamir, M., Moradnejad, P., Nahavandi, S., Khozeimeh, F., Zare, A., Panahiazar, M., Acharya, U. R., & Islam, S. M. S. (2021). Risk factors prediction, clinical outcomes, and mortality in COVID-19 patients. *Journal of Medical Virology*, *93*(4), 2307–2320. https://doi.org/10.1002/jmv.26699

Aljameel, S. S., Khan, I. U., Aslam, N., Aljabri, M., & Alsulmi, E. S. (2021). Machine learning-based model to predict the disease severity and outcome in COVID-19 patients. *Scientific Programming*, *2021*.

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and Unsupervised Learning for Data Science*, 3–21.

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, *9*(7), 1545–1588.

Arti, M. K., & Wilinski, A. (2022). Mathematical modeling and estimation for next wave of COVID-19 in Poland. *Stochastic Environmental Research and Risk Assessment*, *36*(9), 2495–2501.

Ashour, M. A. H. (2022). Optimized Artificial Neural network models to time series. *Baghdad Science Journal*, *19*(4), Article 4. https://doi.org/10.21123/bsj.2022.19.4.0899

Bennett, K. P., & Campbell, C. (2000). Support vector machines: Hype or hallelujah? *ACM SIGKDD Explorations Newsletter*, *2*(2), 1–13.

Bernard, S., Adam, S., & Heutte, L. (2012). Dynamic Random Forests. *Pattern Recognition Letters*, *33*(12), 1580–1586. https://doi.org/10.1016/j.patrec.2012.04.003

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bwire, G. M. (2020). Coronavirus: Why men are more vulnerable to Covid-19 than women? *SN Comprehensive Clinical Medicine*, *2*(7), 874–876.

Chowdhury, M. E. H., Rahman, T., Khandakar, A., Al-Madeed, S., Zughaier, S. M., Doi, S. A. R., Hassen, H., & Islam, M. T. (2021). An Early Warning Tool for Predicting Mortality Risk of COVID-19 Patients Using Machine Learning. *Cognitive Computation*. https://doi.org/10.1007/s12559-020-09812-7

Cockburn, I. M., Henderson, R., & Stern, S. (2019). *The Impact of Artificial Intelligence on Innovation: An Exploratory Analysis. Chap. 4 in The Economics of Artificial Intelligence, edited by AK Agrawal, J. Gans and A. Goldfarb*. University of Chicago Press.

Coombs, C. (2020). Will COVID-19 be the tipping point for the intelligent automation of work? A review of the debate and implications for research. *International Journal of Information Management*, *55*, 102182.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

De Bruyn, A., Verellen, S., Bruckers, L., Geebelen, L., Callebaut, I., De Pauw, I., Stessel, B., & Dubois, J. (2022). Secondary infection in COVID-19 critically ill patients: A retrospective single-center evaluation. *BMC Infectious Diseases*, *22*(1), 207.

Eyre, D. W., Taylor, D., Purver, M., Chapman, D., Fowler, T., Pouwels, K. B., Walker, A. S., & Peto, T. E. (2022). Effect of Covid-19 vaccination on transmission of alpha and delta variants. *New England Journal of Medicine*, *386*(8), 744–756.

Gadekallu, T. R., Rajput, D. S., Reddy, M. P. K., Lakshmanna, K., Bhattacharya, S., Singh, S., Jolfaei, A., & Alazab, M. (2021). A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. *Journal of Real-Time Image Processing*, *18*(4), 1383–1396. https://doi.org/10.1007/s11554-020-00987-8

Gao, H., Zeng, X., & Yao, C. (2019). Application of improved distributed naive Bayesian algorithms in text classification. *The Journal of Supercomputing*, *75*(9), 5831–5847.

Gozes, O., Frid-Adar, M., Greenspan, H., Browning, P. D., Zhang, H., Ji, W., Bernheim, A., & Siegel, E. (2020). Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. *ArXiv Preprint ArXiv:2003.05037*.

Gui, W., Yang, X., Jiang, H., Wu, H., Zeng, M., Wen, Y., Qiu, T., Zhang, Y., Ma, Z., Tong, C., Luo, L., Zhao, Y., & Wang, L. (2021). Prevalence of anxiety and its associated factors among infertile patients after 'two-child' policy in Chongqing, China: A cross-sectional study. *Reproductive Health*, *18*(1), 193. https://doi.org/10.1186/s12978-021-01140-9

Hatmal, M. M., Al-Hatamleh, M. A. I., Olaimat, A. N., Hatmal, M., Alhaj-Qasem, D. M., Olaimat, T. M., & Mohamud, R. (2021). Side Effects and Perceptions Following COVID-19 Vaccination in Jordan: A Randomized, Cross-Sectional Study Implementing Machine Learning for Predicting Severity of Side Effects. *Vaccines*, *9*(6), Article 6. https://doi.org/10.3390/vaccines9060556

Jain, V., & Yuan, J.-M. (2020). Predictive symptoms and comorbidities for severe COVID-19 and intensive care unit admission: A systematic review and meta-analysis. *International Journal of Public Health*, *65*(5), 533–546. https://doi.org/10.1007/s00038-020-01390-7

Jarrar, D. R. (2021). *ASDS7381: SP.TOPIC-ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING*. 17.

Jiang, X., Coffee, M., Bari, A., Wang, J., Jiang, X., Huang, J., Shi, J., Dai, J., Cai, J., & Zhang, T. (2020). Towards an artificial intelligence framework for data-driven prediction of coronavirus clinical severity. *Computers, Materials & Continua*, *63*(1), 537–551.

Kalem, A. K., Kayaaslan, B., Neselioglu, S., Eser, F., Hasanoglu, İ., Aypak, A., Akinci, E., Akca, H. N., Erel, O., & Guner, R. (2021). A useful and sensitive marker in the prediction of COVID-19 and disease severity: Thiol. *Free Radical Biology and Medicine*, *166*, 11–17. https://doi.org/10.1016/j.freeradbiomed.2021.02.009

Laatifi, M., Douzi, S., Bouklouz, A., Ezzine, H., Jaafari, J., Zaid, Y., El Ouahidi, B., & Naciri, M. (2022). Machine learning approaches in Covid-19 severity risk prediction in Morocco. *Journal of Big Data*, *9*(1), 5. https://doi.org/10.1186/s40537-021-00557-0

Lei, J., Li, M., & Wang, X. (2021). Predicting the development trend of the second wave of COVID-19 in five European countries. *Journal of Medical Virology*, *93*(10), 5896–5907.

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. *International Conference on Information Computing and Applications*, 246–252.

Mancilla-Galindo, J., Kammar-García, A., Martínez-Esteban, A., Meza-Comparán, H. D., Mancilla-Ramírez, J., & Galindo-Sevilla, N. (2021). COVID-19 patients with increasing age experience differential time to initial medical care and severity of symptoms. *Epidemiology & Infection*, *149*, e230. https://doi.org/10.1017/S095026882100234X

Meyer, D. (2015). *Support Vector Machines ∗ The Interface to libsvm in package e1071*. C). For more information, see: http://www.csie.ntu.edu.tw/~cjlin/papers/ijcnn.ps.gz.

Naik, N., Hameed, B. M., Shetty, D. K., Swain, D., Shah, M., Paul, R., ... & Somani, B. K. (2022). Legal and ethical consideration in artificial intelligence in healthcare: who takes responsibility?. *Frontiers in surgery*, 9, 266.

Nemati, M., Ansary, J., & Nemati, N. (2020). Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data. *Patterns*, *1*(5), 100074. https://doi.org/10.1016/j.patter.2020.100074

Nordström, P., Ballin, M., & Nordström, A. (2022). Risk of infection, hospitalisation, and death up to 9 months after a second dose of COVID-19 vaccine: A retrospective, total population cohort study in Sweden. *The Lancet*, *399*(10327), 814–823.

Ocak, H. (2013). A medical decision support system based on support vector machines and the genetic algorithm for the evaluation of fetal well-being. *Journal of Medical Systems*, *37*(2), 9913. https://doi.org/10.1007/s10916-012-9913-4

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *ArXiv Preprint ArXiv:1511.08458*.

Prem, K., Liu, Y., Russell, T. W., Kucharski, A. J., Eggo, R. M., Davies, N., Flasche, S., Clifford, S., Pearson, C. A. B., Munday, J. D., Abbott, S., Gibbs, H., Rosello, A., Quilty, B. J., Jombart, T., Sun, F., Diamond, C., Gimma, A., Zandvoort, K. van, … Klepac, P. (2020). The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study. *The Lancet Public Health*, *5*(5), e261–e270. https://doi.org/10.1016/S2468-2667(20)30073-6

Raschka, S. (2014). Naive bayes and text classification i-introduction and theory. *ArXiv Preprint ArXiv:1410.5329*.

Ray, S. (2019). A quick review of machine learning algorithms. *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 35–39.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review,* 65(6), 386.

Sachs, J. D., Karim, S. S. A., Aknin, L., Allen, J., Brosbøl, K., Colombo, F., Barron, G. C., Espinosa, M. F., Gaspar, V., & Gaviria, A. (2022). The Lancet Commission on lessons for the future from the COVID-19 pandemic. *The Lancet*, *400*(10359), 1224–1280.

Salman, A. D. (2020). Study impact the latitude on Covid-19 spread virus by data mining algorithm. *Study Impact the Latitude on Covid-19 Spread Virus by Data Mining Algorithm*, *1664*(12), 109–120.

Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 5–22). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_1

Sanz-Muñoz, I., Tamames-Gómez, S., Castrodeza-Sanz, J., Eiros-Bouza, J. M., & de Lejarazu-Leonardo, R. O. (2021). Social distancing, lockdown and the wide use of mask; a magic solution or a double-edged sword for respiratory viruses epidemiology? *Vaccines*, *9*(6), 595.

Shadeed, S., & Alawna, S. (2021). GIS-based COVID-19 vulnerability mapping in the West Bank, Palestine. *International Journal of Disaster Risk Reduction*, *64*, 102483. https://doi.org/10.1016/j.ijdrr.2021.102483

Song, J., Xie, H., Gao, B., Zhong, Y., Gu, C., & Choi, K.-S. (2021). Maximum likelihood-based extended Kalman filter for COVID-19 prediction. *Chaos, Solitons & Fractals*, *146*, 110922.

Vigón, L., Fuertes, D., García-Pérez, J., Torres, M., Rodríguez-Mora, S., Mateos, E., Corona, M., Saez-Marín, A. J., Malo, R., Navarro, C., Murciano-Antón, M. A., Cervero, M., Alcamí, J., García-Gutiérrez, V., Planelles, V., López-Huertas, M. R., & Coiras, M. (2021). Impaired Cytotoxic Response in PBMCs From Patients With COVID-19 Admitted to the ICU: Biomarkers to Predict Disease Severity. *Frontiers in Immunology*, *12*. https://www.frontiersin.org/articles/10.3389/fimmu.2021.665329

Viloria Winnett, A., Akana, R., Shelby, N., Davich, H., Caldera, S., Yamada, T., Reyna, J. R. B., Romano, A. E., Carter, A. M., & Kim, M. K. (2022). Extreme differences in SARS-CoV-2 Omicron viral loads among specimen types drives poor performance of nasal rapid antigen tests for detecting presumably pre-infectious and infectious individuals, predicting improved performance of combination specimen antigen tests. *MedRxiv*, 2022.07. 13.22277513.

Wake, R. M., Morgan, M., Choi, J., & Winn, S. (2020). Reducing nosocomial transmission of COVID-19: Implementation of a COVID-19 triage system. *Clin Med (Lond)*, *20*(5), e141–e145.

Wang, G., Zhang, Q., Zhao, X., Dong, H., Wu, C., Wu, F., Yu, B., Lv, J., Zhang, S., Wu, G., Wu, S., Wang, X., Wu, Y., & Zhong, Y. (2020). Low high-density lipoprotein level is correlated with the severity of COVID-19 patients: An observational study. *Lipids in Health and Disease*, *19*(1), 204. https://doi.org/10.1186/s12944-020-01382-9

Wang, L., & Cheng, G. (2022). Sequence analysis of the emerging SARS-CoV-2 variant Omicron in South Africa. *Journal of Medical Virology*, *94*(4), 1728–1733.

Wang, S.-C. (2003). Artificial Neural Network. In S.-C. Wang, *Interdisciplinary Computing in Java Programming* (pp. 81–100). Springer US. https://doi.org/10.1007/978-1-4615-0377-4_5

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C., & Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), Article 7798. https://doi.org/10.1038/s41586-020-2008-3

Xiao, F., Sun, R., Sun, W., Xu, D., Lan, L., Li, H., Liu, H., & Xu, H. (2021). Radiomics analysis of chest CT to predict the overall survival for the severe patients of COVID-19 pneumonia. *Physics in Medicine & Biology*, *66*(10), 105008. https://doi.org/10.1088/1361-6560/abf717

Xiong, D., Zhang, L., Watson, G. L., Sundin, P., Bufford, T., Zoller, J. A., Shamshoian, J., Suchard, M. A., & Ramirez, C. M. (2020). Pseudo-likelihood based logistic regression for estimating COVID-19 infection and case fatality rates by gender, race, and age in California. *Epidemics*, *33*, 100418. https://doi.org/10.1016/j.epidem.2020.100418

Xiong, Y., Ma, Y., Ruan, L., Li, D., Lu, C., Huang, L., & the National Traditional Chinese Medicine Medical Team. (2022). Comparing different machine learning techniques for predicting COVID-19 severity. *Infectious Diseases of Poverty*, *11*(1), 19. https://doi.org/10.1186/s40249-022-00946-4

Yan, L., Zhang, H.-T., Xiao, Y., Wang, M., Guo, Y., Sun, C., Tang, X., Jing, L., Li, S., Zhang, M., Xiao, Y., Cao, H., Chen, Y., Ren, T., Jin, J., Wang, F., Xiao, Y., Huang, S., Tan, X., … Yuan, Y. (2020). Prediction of criticality in patients with severe Covid-19 infection using three clinical features: A machine learning-based prognostic model with clinical data in Wuhan. *MedRxiv*. https://doi.org/10.1101/2020.02.27.20028027

Yang, C., & Wang, J. (2021). Modeling the transmission of COVID-19 in the US–A case study. *Infectious Disease Modelling*, *6*, 195–211.

Supplements

Appendix(A) Permission letter

جامعة بيرزيت
BIRZEIT UNIVERSITY
Office of the President مكتب رئيس الجامعة

4 تشرين الثاني 2021

معالي الدكتورة مي الكيلة المحترمة
وزيرة الصحة

تحية طيبة وبعد،

الموضوع: طلب تسهيل مهمة لطالبة في برنامج ماجستير الإحصاء التطبيقي وعلم البيانات

أطيب الأمنيات نرسلها لكم من جامعة بيرزيت بدوام الصحة والعافية، وبالإشارة إلى الموضوع أعلاه نرجو التكرم بتسهيل مهمة الطالبة أسماء محمد شعيبي (1205039) من برنامج ماجستير الإحصاء التطبيقي وعلم البيانات، في بحثها: **"التنبؤ بوتيرة الموجات القادمة لفايروس كورونا (كوفيد -19) باستخدام خوارزميات الذكاء الصناعي"**. والذي يتطلب جمع بيانات من وزارة الصحة.

شاكرين لكم تعاونكم الدائم مع جامعة بيرزيت.

وتفضلوا بقبول فائق الاحترام والتقدير ،،،

د. بشارة دوماني
رئيس الجامعة

- نسخة/ نائب الرئيس للشؤون الأكاديمية
/ عميد الدراسات العليا
/ الطالبة أسماء شعيبي

P.O.Box 14 Birzeit, Palestine, Tel +970 2 298 2008, Fax +970 2 281 0656 ص ب 14 بيرزيت، فلسطين، هاتف 2008 298 2 970+، فاكس 0656 281 2 970+
president@birzeit.edu www.birzeit.edu

Appendix(B)) R Codes

#SVM (SEVERITY(1)_NON LINEAR)

```
install.packages("caret")
library("caret")
attach(icu_num)
data <- icu_num
str(data)
head(str)
set.seed(3033)
intrain <- createDataPartition(y=data$severity, p = 0.7, list = FALSE)
trainig <- data[intrain,]
testing <- data [-intrain,]
dim(trainig)
dim(testing)
anyNA(data)
summary(data)
trainig[["severity"]] = factor(trainig[["severity"]])
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
set.seed(3233)
library(e1071)
model <- svm(severity ~ ., data = data)
svm_linear <- train(severity~., data = trainig, method = "svmLinear",
 trControl = trctrl,
 preProcess = c("center", "scale"),
```

```
 tuneLength = 10)
#confusion matrix and prediction
test_pred <- predict(svm_linear, newdata = testing)
test_pred
confusionMatrix(table(test_pred, testing$severity))
grid <- expand.grid(C = c(0,0.01,0.05,0.1,0.25,0.5,0.75,1,1.25,1.5,1.75,2.5))
svm_linear_Grid <- train(severity~., data = trainig, method = "svmLinear",
 preProcess = c("center", "scale"),
 tuneGrid = grid,
 tuneLength = 10)
svm_linear_Grid
plot(svm_linear_Grid)
test_pred_grid <- predict(svm_linear_Grid, newdata = testing)
test_pred_grid
confusionMatrix(table(test_pred_grid, testing$severity))
```

##SVM (SEVERITY(2)_NON LINEAR) ALL DATA IN ICU

```
install.packages("caret")
library("caret")
library(ggplot2)
library(lattice)
install.packages("ggplot2")
attach(icu_for_r)
data <- icu_for_r
str(data)
head(str)
set.seed(3033)
intrain <- createDataPartition(y=data$objective, p = 0.8, list = FALSE)
```

```
trainig <- data[intrain,]
testing <- data [-intrain,]
dim(trainig)
dim(testing)
anyNA(data)
summary(data)
trainig[["objective"]] = factor(trainig[["objective"]])
trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
set.seed(3233)
library(e1071)

model <- svm(objective ~ ., data = data)
svm_linear <- train(objective~., data = trainig, method = "svmLinear",
                    trControl = trctrl,
                    preProcess = c("center", "scale"),
                    tuneLength = 10)
#confusion matrix and prediction
test_pred <- predict(svm_linear, newdata = testing)
test_pred
confusionMatrix(table(test_pred, testing$objective))
grid <- expand.grid(C = c(0,0.01,0.05,0.1,0.25,0.5,0.75,1,1.25,1.5,1.75,2.5))
svm_linear_Grid <- train(objective~., data = trainig, method = "svmLinear",
                         preProcess = c("center", "scale"),
                         tuneGrid = grid,
                         tuneLength = 10)
svm_linear_Grid
plot(svm_linear_Grid)
```

```
test_pred_grid <- predict(svm_linear_Grid, newdata = testing)
test_pred_grid
confusionMatrix(table(test_pred_grid, testing$objective))
#####
qplot(data$Chronicdiseases, data$Diagnosis, data = data,
      color = data$objective)
mymodel <- svm(objective~., data = data)
summary(mymodel)
plot(mymodel, data = data,
     hospital~Chronicdiseases,
     slice = list(Result = 3, EOSINPHILS = 4))
set.seed(123)
tmodel <- tune(svm, objective~., data= data,
               ranges = list(epsilon = seq(0,1,0.1), cost = 2^(2:9)))
plot(tmodel)
win.graph(12,4,12)
plot(tmodel)
```

```
#RANDOM FOREST
library(randomForest)
library(datasets)
library(caret)
library(lattice)
data<-icu_num_all
str(data)
data$severity <- as.factor(data$severity)
```

```
table(data$severity)

set.seed(222)

ind <- sample(2, nrow(data), replace = TRUE, prob = c(0.7, 0.3))

train <- data[ind==1,]

test <- data[ind==2,]

#Random Forest

rf <- randomForest(severity~., data=train, proximity=TRUE)

print(rf)


randomForest(formula = severity ~ ., data = train)

p1 <- predict(rf, train)

confusionMatrix(p1, train$severity)

p2 <- predict(rf, test)

confusionMatrix(p2, test$severity)

plot(rf)

#Type of Random Forest: classification


confusionMatrix(table(p2, test$severity))

cm <- confusionMatrix(p1, train$severity)

cm$sever[3,1]

# Tune mtry

library(randomForest)

model_tuned<-tuneRF(x=data[,2:19],y=data$severity,ntreeTry = 3000,mtryStart = 3,stepFactor
= 1.5,improve = 0.01,trace = TRUE)

win.graph(20,12,12)

model_tuned
```

```
# No. of nodes for the trees
hist(treesize(rf),
 main = “No. of Nodes for the Trees”,
 col = “green”)
# Variable Importance
varImpPlot(rf,
 sort = T,
 n.var = 10,
 main = “Top 10 - Variable Importance”)
importance(rf)
varUsed(rf)
# Partial Dependence Plot
partialPlot(rf, train, severity, “2”)
# Extract Single Tree
getTree(rf, 1, labelVar = TRUE)
# Multi-dimensional Scaling Plot of Proximity Matrix
MDSplot(rf, train$severity)
```