DATA MINING APPROACH
TO CLASSIFY COVID-19 SEVERITY BY CLINICAL SYMPTOMS

LAURA JASMINE ANAK THOMAS KANYAN

UNIVERSITI SAINS MALAYSIA

July 2021

DATA MINING APPROACH TO CLASSIFY COVID-19 SEVERITY BY CLINICAL SYMPTOMS

By:

LAURA JASMINE ANAK THOMAS KANYAN
(Matrix no: 138286)

Supervisor:

Assoc. Prof. Dr Loh Wei Ping

July 2021

This dissertation is submitted to
Universiti Sains Malaysia
As partial fulfilment of the requirement to graduate with honours degree in
BACHELOR OF ENGINEERING (MANUFACTURING ENGINEERING WITH MANAGEMENT)

School of Mechanical Engineering
Engineering Campus
Universiti Sains Malaysia

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed……………………………………………… (Laura Jasmine Anak Thomas Kanyan)

Date…………………………………………… (12/7/2021)

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated.

Other sources are acknowledged by giving explicit references.

Bibliography/references are appended.

Signed……………………………………………… (Laura Jasmine Anak Thomas Kanyan)

Date…………………………………………… (12/7/2021)

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for interlibrary loan, and for the title and summary to be made available outside organizations.

Signed……………………………………………… (Laura Jasmine Anak Thomas Kanyan)

Date…………………………………………… (12/7/2021)

ACKNOWLEDGEMENT

I would like to express my utmost gratitude to my supervisor, Associate Professor Dr Loh Wei Ping for her guidance throughout this research study. Throughout this project, she has given me unwavering support and useful advice as well. Thanks to her, I have been able to gain more understanding on the techniques used in data mining through her data mining lectures. Besides that, this project itself, to me, was a very beneficial project because it provided me the awareness regarding COVID-19 severity. To be given the opportunity to study more about this disease would not be possible without her. Furthermore, I truly think that my skills in data analysis and in data interpretation has grown thanks to my supervisor as well. These skillsets gained from this project, I believe, will benefit me greatly in my future career.

Besides that, I would like to thank my family for their support and care throughout this pandemic. They too, have encouraged me to do my best and ensure that I am able to complete this project in time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACE-2	Angiotensin-converting-enzyme 2
ARDS	Acute Respiratory Distress Syndrome
AUC	Area Under Curve
CD4	Cluster of Differentiation 4
CDC	Centres for Disease Control and Prevention
CLD	Chronic Liver Disease
COPD	Chronic Obstructive Pulmonary Disorder
COVID-19	Coronavirus Disease 2019
CSV	Comma Separated Values
CT	Computerized Tomography
CVD	Cerebrovascular Disease
ENT	Ear Nose Throat
IL-6	Interleukin-6
MCO	Movement Control Order
NLR	Neutrophil-Lymphocyte Ratio
OFS	Olfactory Score
RT-PCR-CT	Reverse transcription polymerase chain reaction (Cycle Threshold)
SMO	Sequential Minimal Optimization
STL	Smell and/or loss of sense of taste
SVM	Support Vector Machine
VAS	Visual Analogue Scale
WEKA	Waikato Environment for Knowledge Analysis
WHO	World Health Organization

LIST OF APPENDICES

ABSTRAK

Penyakit Coronavirus (COVID-19) merupakan sebuah kebimbangan global memandangkan ia merebak ke seluruh dunia, menjangkiti jutaan manusia. Mereka yang dijangkiti menunjukkan simtom seperti demam, batuk, keletihan, sakit kepala, sesak nafas, sakit tekak, mialgia, artralgia, loya/mual, cirit-birit, sakit dada, hilang deria bau dan rasa. Walaupun banyak kajian telah dijalankan berkenaan penyakit ini, hubungan antara simtom dan keparahan penyakit masih tidak jelas. Beberapa kajian menggunakan pendekatan perlombongan data dalam mengkelaskan tahap keparahan COVID-19 berdasarkan simtom. Oleh itu, matlamat kajian ini bukan sahaja untuk menentukan petunjuk keparahan COVID-19 tapi juga untuk mengenalpasti simtom awal COVID-19 dan untuk membina model yang mengkaitkan hubungan antara simtom COVID-19 untuk meramal tahap keparahan COVID-19. Perisian yang digunakan dalam kajian ini untuk analisis perlombongan data merupakan Waikato Environment for Knowledge Analysis (WEKA) versi 3.8. Pengumpulan data yang melibatkan dua kajian kes berkaitan dengan COVID-19 telah diambil dari Kaggle dan sebuah jurnal kajian dari Turki. Pra-pemprosesan data telah dijalankan untuk mengenal pasti dan membuang unsur luaran. Nilai yang hilang telah dirawat menggunakan kaedah penapisan dan melalui pengisian nilai. Algoritma pengelasan: J48, SMO, Random Forest, dan Simple Logistic telah dijalankan dan diuji untuk mengkelaskan data kepada tiga kelas: ringan, sederhana, dan teruk. Hasil menunjukkan bahawa simtom seperti dispnea dan kesukaran pernafasan adalah penunjuk utama untuk keparahan COVID sementara mereka yang mengalami simtom seperti hilang deria bau lebih cenderung untuk dikategorikan sebagai tahap COVID yang ringan atau sederhana.

DATA MINING APPROACH TO CLASSIFY COVID-19 SEVERITY BY CLINICAL SYMPTOMS

ABSTRACT

Coronavirus Disease (COVID-19) is a global concern as it has spread throughout the world, infecting millions. Those infected were presented with symptoms like fever, cough, fatigue, headache, shortness of breath, sore throat, myalgia, arthralgia, nausea, diarrhoea, chest pain, loss of smell and taste. Although there have been studies carried out regarding this disease, the relationship between the symptoms and the disease severity remains unclear. Few studies have used data mining approaches in classifying COVID-19 severity levels based on symptoms. Therefore, the goal of this study was not only to determine the severity indicators of COVID-19 but also to identify early symptoms of COVID-19 and to model the relationship between COVID-19 symptoms to predict COVID-19 severity levels. The software used in this study for data mining analysis was the Waikato Environment for Knowledge Analysis (WEKA) version 3.8. The data collection which involves two case studies related to COVID-19 were retrieved from Kaggle and a research journal from Turkey. Data pre-processing was carried out to identify and remove outliers. Missing values were treated using filtering and imputation methods. The classification algorithms: J48, SMO, Random Forest, and Simple Logistic were executed and tested to classify data into three classes: mild, moderate, and severe. Results show that symptoms like dyspnoea and breathing problems were the main indicators of severe COVID while those experiencing symptoms like loss of smell were more likely to be categorized under mild or moderate COVID level.

CHAPTER 1 : INTRODUCTION

1.1 Overview

This chapter provides an introduction on the study of COVID-19 symptoms severity. The symptoms of COVID-19, the disease severity as well as its characteristics, and the high-risk groups of COVID-19 were highlighted. The project background of this research will be presented. Next, the list of objectives that are to be accomplished are highlighted. Then, the knowledge gap and the scope of work for this research study are outlined.

1.2 Coronavirus Disease (COVID-19)

COVID-19 or Coronavirus disease 2019 is a virus that has spread globally. Those infected with this virus were reportedly experiencing symptoms like fever, cough, fatigue as well as other discomforts like sore throat, diarrhoea, headache, nausea, myalgia, chest pain and loss of smell and taste. As this disease is a novel virus, research works have been carried to study its severity and symptoms. Past studies have also looked into clinical characteristics, epidemiological characteristics, laboratory findings, and radiographic findings of the disease.

Although COVID-19 can be symptomatic, it can also be symptomless and yet contagious. There are cases where people would experience worsened symptoms like shortness of breath and pneumonia a week after the onset of symptoms. The elderly and those who have existing chronic medical conditions commonly known to have a higher risk of developing severe COVID symptoms. Furthermore, smoking and obesity are among the lifestyle factors that could contribute to the severity of the disease. High-risk individuals can also experience complications like pneumonia, acute respiratory distress syndrome (ARDS), acute kidney injury, blood clots, heart diseases and organ failure in several organs.

The severity of the disease can be categorized from mild to severe levels depending on the progression of disease symptoms. Patients classified as mild were those who showed symptoms like fever, cough, sore throat, malaise, headache, muscle pain, nausea, vomiting, diarrhea, loss of taste and smell but no shortness of breath, dyspnoea, or abnormal imaging. According to the World Health Organization (WHO, 2020), a severe case is defined as

1) Shortness of breath, respiratory rate > 30 breaths/min;

2) Oxygen saturation at rest < 93%; or

3) Partial pressure of arterial oxygen (PaO2)/fraction of inspiration oxygen (FiO2) ⩽ 300 mmHg; or

4) A requirement of mechanical ventilation

While the resource requirements for those in severe cases requires mechanical ventilation and oxygen therapy, those who experience only mild or moderate symptoms would be placed in isolation. Nevertheless, current advancement in technology has helped in keeping the disease under control, from real-time data collection via smartphone applications for tracking of the disease, to mass production of test swabs that can allow for quick detection of infections.

1.3 Project Background

Malaysia reported its first COVID-19 case on the 25th January 2020 (New Straits Times, 2020). The disease spreads so rapidly that a Movement Control Order (MCO) was imposed by the government at two levels: MCO 1.0 from 18 March 2020 to 31 March followed by extensions from 1 April till 12 May (phases 2,3, and 4) for prevention and control of the disease. All government and private offices, schools, kindergarten, and higher education institutions which includes both private and government sectors were temporarily closed. During the MCO period, Malaysians were banned from travelling overseas while those returning from overseas will have to go through health checks as well as self-quarantine for 14 days.

Early diagnosis of the virus is performed via real-time quantitative polymerase chain reaction (rt-PCR) in throat swabs. Among the commonly reported symptoms of COVID-19 include fever, cough, and fatigue whereas the less common ones are sore throat, headache, loss of smell and taste as well as gastrointestinal symptoms like nausea and diarrhoea. Symptoms like chest tightness, dyspnoea, or shortness of breath could indicate a severe case. The disease severity can range from mild or asymptomatic to severe and critically severe. In fact, it is important to understand the severity classification as it helps in determining the disease progression. Early identification of factors that contribute to severity can help in preventing unfavourable outcomes as these are markers before the cases become severe. On the other hand, there are also asymptomatic cases whereby patients do not exhibit symptoms but are positive COVID-19 carriers. Asymptomatic carriers are usually categorized under the mild symptom category. Besides, those who are at high risk of developing severe symptoms include the elderly and those with underlying diseases like asthma, hypertension, diabetes, pulmonary disease, cardiovascular disease, cancer, liver, or kidney disease, stroke, and dementia.

Hence, a relationship between the multiple factors consisting of symptoms, risk factors, laboratory findings, and radiographic findings can give an early prediction on the COVID-19 severity levels.

1.4 Objectives

The objectives of this project are:

a) to identify the early symptoms of COVID-19

b) to determine the severity indicators of COVID-19

c) to model the relationship between COVID-19 symptoms to predict COVID-19 severity levels

1.5 Problem Statement

The common symptoms of COVID-19 vary from one person to another, ranging from mild to critically severe. While there were many ongoing research studies, the relationship of the developed symptoms is still uncertain. The severity of the COVID-19 symptoms has been reported based on multiple aspects. But little is known on how these symptoms are related to one another. Most studies had a limited or small number of cases. Besides, there was incomplete documentation regarding clinical symptoms and laboratory findings. Moreover, most studies were mainly retrospective. Confounding errors and biasness were common in retrospective studies.

1.6 Scope of Work

Due to the limited publicly available data regarding symptoms of COVID-19 for Malaysian cases, this study would focus on available datasets globally. The global datasets were retrieved from Kaggle data repository [71] and a dataset from Turkey by Salepci et al. [72]. The study attributes considered were age, gender, symptoms: fever, cough, sore throat, headache, myalgia, arthralgia, dyspnoea, loss of smell and taste, fatigue, chest pain, rhinorrhea, inappetence and underlying medical health conditions: high blood pressure, diabetes, heart disease, lung disease, stroke. Multiple datasets were examined, patterns and correlation between symptoms are discovered in addition to identifying the independent (age, gender, COVID-19 symptoms, comorbidities) and dependent variables (patients' risk level and patient clinical status) to provide insight into COVID-19 severity. Data analysis were conducted using data mining approach at four levels: data pre-processing stage, classification analysis, model building, and model verification. The software used for data mining analysis was the Waikato Environment for Knowledge Analysis (WEKA) version 3.8.

1.7 Outline

This thesis is divided into 5 chapters.

Chapter 1 provides an introduction on the Coronavirus disease (COVID-19) and its symptoms in addition to its severity levels. The project background, objectives, problem statement and scope of work are discussed. The main focus of this chapter is to present the background of COVID-19 disease, reported symptoms, the objectives and scope of the research study.

Chapter 2 presents the literature review on COVID-19 symptoms as well as the factors affecting the severity of the disease. This topic mainly contains the review of past studies related to COVID-19 symptoms and its severity. The findings of the previous papers are also highlighted.

Chapter 3 provides an outline of the methodology carried out for this project. This chapter explains the methods used in beginning from the data collection stage followed by the data pre-processing stage and the classification analysis. The data attributes used for classification are further discussed.

The results that are obtained from this study are presented in Chapter 4. The results consist of the attributes used to build the classification model in WEKA and the prediction accuracies obtained by the predictive models. The classification outcomes are further studied, analysed, and compared. The main attributes especially the symptoms which distinguish the severity levels of COVID are highlighted.

Chapter 5 summarizes and conclude the whole research study. The conclusion is based on the results obtained in Chapter 4. Limitations and the potential future works are mentioned in this chapter.

CHAPTER 2 : LITERATURE REVIEW

2.1 Overview

COVID-19, a disease caused by the SARS-CoV-2 infection, has affected the lives of many people. The disease severity can range from mild or asymptomatic to severe and critically severe. Symptoms may appear from two to 14 days after exposure to the virus. Health complications due to COVID-19 includes mild and severe pneumonia, acute respiratory distress syndrome, sepsis, and septic shock (WHO,2020). The diagnosis of COVID-19 is similar to other viral infections, by extracting blood, saliva, and tissue samples. This chapter presents the state-of-the-art review of the signs and symptoms of the disease, the factors affecting the severity of COVID-19. There were altogether 53 related works retrieved from a single database of year 2019 till 2020.

2.2 Search Strategy

Articles were sought using the Science Direct database (Figure 2.1). The basic keywords used for the search was 'COVID-19 symptoms' filtered by year 2019 and 2020. This is because COVID-19 was only discovered in 2019. The number of articles related to the search was 12447. Subsequent filter was based on the categories in which only research articles were considered. This resulted in 4599 relevant articles. Next, the articles were selected if the titles or the abstracts fulfil any of the following conditions:

1) Contain the keyword 'symptoms'. If the abstract of the articles contains the keyword 'symptoms', then the articles would be included.

2) Described the epidemiology (prevalence), clinical characteristics, or clinical manifestations of COVID-19.

3) Listed the symptoms of COVID-19 (e.g. fever, cough, gastrointestinal symptoms like nausea and diarrhoea, olfactory dysfunctions like loss of smell and/or

taste, etc.). This means that if the abstract contains a list of COVID-19 related symptoms, the articles would be selected.

4) Involving symptomatic AND/OR asymptomatic cases.

After screening based on titles and abstracts, 4482 records were excluded, and 117 articles were left (Figure 2.1). A total of 56 articles were removed as they were pre-proof journals while the remaining 61 full-text articles were then assessed for eligibility based on the following criteria:

Inclusion criteria:

- Written in English.
- Retrospective studies, retrospective case-control studies, retrospective cohort studies, cross-sectional study, prospective cohort studies, and others (e.g., using big Data)

Exclusion criteria:

- Written in a non-English language.

Out of the 61 full-text articles, 8 records were excluded because the articles belong to review type. Thus, the remaining 53 articles were included for review.

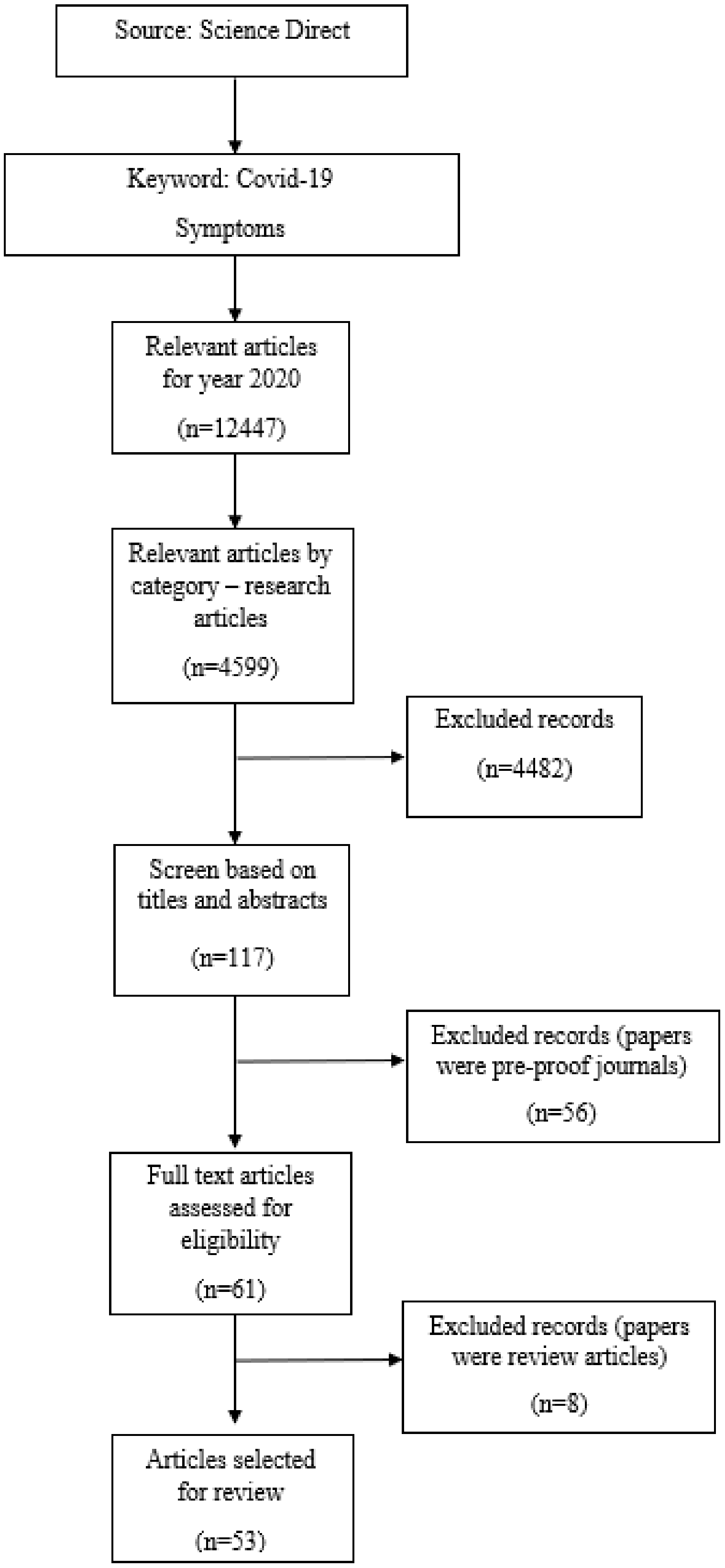

Source: Science Direct
Keyword: Covid-19 Symptoms
Relevant articles for year 2020 (n=12447)
Relevant articles by category – research articles (n=4599)
Excluded records (n=4482)
Screen based on titles and abstracts (n=117)
Excluded records (papers were pre-proof journals) (n=56)
Full text articles assessed for eligibility (n=61)
Excluded records (papers were review articles) (n=8)
Articles selected for review (n=53)

Figure 2.1: Schematic diagram of article search strategy from the ScienceDirect platform

2.3 COVID-19

COVID-19 or Coronavirus disease 2019 is a virus that has spread globally. Those infected with this virus were reportedly experiencing symptoms like fever, cough, fatigue as well as other discomforts like sore throat, diarrhoea, headache, headache, nausea, myalgia, chest pain and loss of smell and taste. Those exposed to COVID-19 virus may exhibit symptoms or no symptoms (asymptomatic). There are cases where people would experience worsened symptoms such like shortness of breath and pneumonia a week after the onset of symptoms. The elderly and those who have existing chronic medical conditions have a higher risk of developing severe symptoms [77], [78]. Health complications due to COVID-19 includes mild and severe pneumonia, acute respiratory distress syndrome, sepsis, and septic shock (WHO,2020). As the disease is relatively new, various research works have been carried out to study its severity and the disease distribution. Past studies have considered the clinical characteristics, epidemiological characteristics, laboratory findings, and radiographic findings of the disease.

2.4 Symptoms of COVID-19

Many papers have reported on the epidemiological characteristics and clinical characteristics that can be linked to disease severity. Reports of these characteristics were first discovered in China [1], [2]. Clinical data were collected at the early stage in South Korea which consisted of all cases, including mild, severe, and asymptomatic cases [3]. In Japan, reports highlighted the clinical features of people who were infected with COVID-19 with various degrees of severity showing that serum LDH, age, consolidation on CT scan, and lymphopenia can be the predictors of COVID disease severity [4]. Countries in the Middle East like Jordan [5] and Iraq [6] had mostly patients who were reported either mild or asymptomatic. In Iran, cases were mostly either with no symptoms or have mild symptoms [7]. Reports in Oman also helped to shed light on risk factors and outcomes of their hospitalized patients [8]. Brendish et al. [9], whereas, showed that there were significant differences regarding the clinical characteristics, symptoms, and clinical outcomes in both positive and negative COVID-19 patients.

Multiple studies have reported symptoms of COVID-19 and the severity of the symptoms such as fever, cough, and fatigue as the prevalent symptoms, while other symptoms also include sore throat, myalgia, diarrhoea, nausea, loss of smell and taste as well as shortness of breath and chest tightness (Figure 2.2 and Table 2.1). Figure 2.2 shows the statistics of related studies from year 2019 to 2020. Only 53 papers were considered as these papers contains symptoms data. More than half of the papers reported of symptoms like fever, cough, and fatigue. Less than 10 papers reported lack of appetite, chest pain and runny nose. Table 2.1 presents the papers which reported the COVID-19 symptoms. There are in total of 13 symptoms that were reported in previous studies.

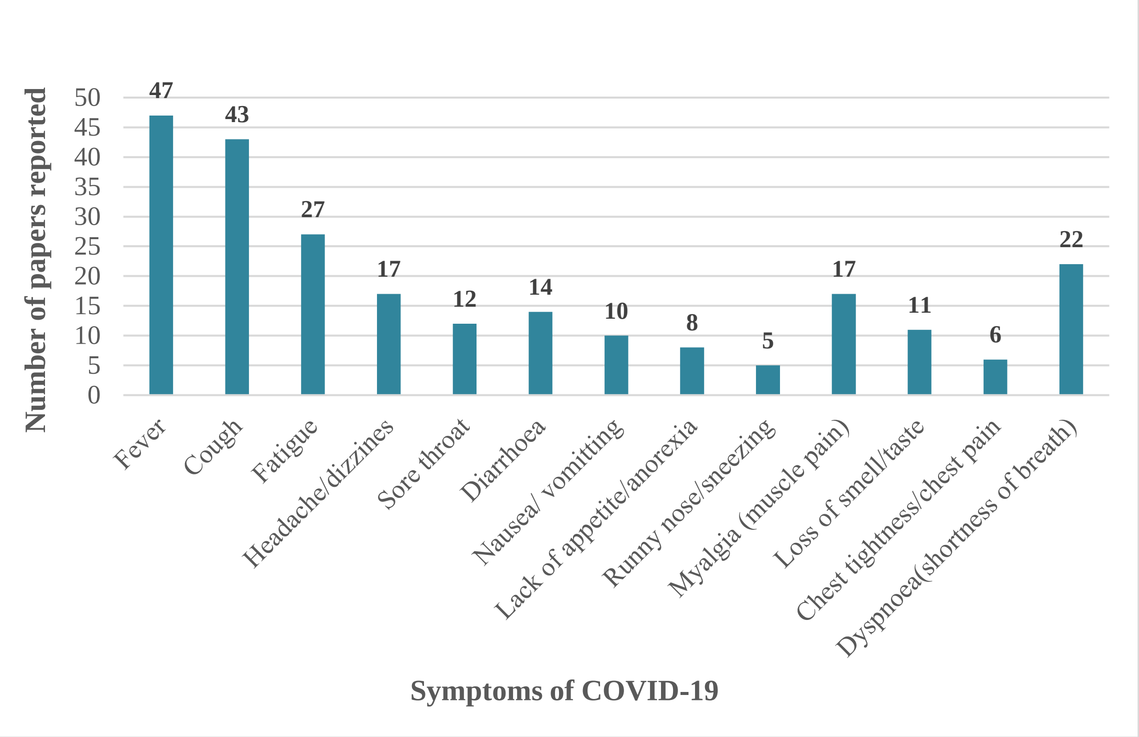


Figure 2.2: Symptoms reported by number of research articles out of 53 papers reviewed

Table 2.1: Papers that reported COVID-19 symptoms

COVID-19 Symptoms	**References**
Fever	[1], [2], [11]–[20], [3], [21]–[30], [4], [31]–[40], [5], [41]–[47], [6]–[10]
Cough	[1], [2], [12]–[15], [17]–[22], [3], [23], [24], [26]–[33], [4], [34]–[36], [38]–[41], [46]–[48], [6], [49], [7]–[11]
Fatigue	[5], [6], [21], [25], [26], [28], [30], [33], [34], [36], [38], [39], [7], [40]–[42], [44], [46], [48], [49], [9], [11], [14], [17]–[20]
Headache/dizziness	[3], [6], [42], [44], [48], [49], [11], [15], [21], [23], [27], [34], [36], [40]
Sore throat	[2], [5], [39], [40], [7], [10], [15], [19], [23], [30], [34], [38]
Diarrhoea	[3], [14], [37], [39], [40], [49], [15], [17], [18], [22], [25], [27], [32], [34]
Nausea and/or vomiting	[7], [8], [17], [21], [22], [25], [34], [45], [48], [49]
Lack of appetite/anorexia	[5], [7], [9], [19], [22], [25], [32], [37]
Runny nose/sneezing	[2], [3], [5], [13], [42]

Myalgia	[3], [6], [28], [32], [34], [36], [38], [39], [49], [7], [11], [17], [18], [23], [25]–[27]
Loss of smell and/or taste	[3], [6], [42], [50], [11]–[13], [16], [19], [29], [32], [33]
Chest tightness/chest pain	[2], [5], [9], [17], [26], [32]
Shortness of breath/dyspnoea	[1], [2], [26], [29], [31]–[34], [36], [37], [39], [40], [5], [45], [46], [10], [11], [14], [15], [17], [18], [20]

Alsofayan et al. [10] and Yoshimura et al. [11] showed that fever and cough were the most common symptoms. Even in other diseases like severe acute respiratory syndrome (SARS) and Middle East Respiratory Syndrome (MERS), fever and cough would be the initial symptoms presented [12]. Fever can be further categorized into mild (less than 37.3 ºC), moderate (37.3 – 38.0 ºC), and high (above 38.0ºC) based on the Centers for Disease Control and Prevention (CDC) guidelines. High fever does not always equate to severe infection of the disease. A study by Li et al. [13] showed that the group which is in the mild category for severity had a mean temperature of 38.0°C whereas the group in the normal category had a mean temperature of 38.0ºC (range 37.8-38.5ºC). The severe group, on the other hand, was 38.4ºC (range 38.0-38.9ºC). Their findings showed that fever was not associated with disease severity. Although fever has always been a dominant symptom, there were cases whereby cough symptoms were highly diagnosed [14].

According to Shah et al. [15], COVID-19 patients who experience a longer duration of symptoms, particularly fatigue, fever, and myalgia would be admitted in the hospital for a longer period. Moreover, this group was more likely to develop acute respiratory distress syndrome (ARDS). On top of clinical features, laboratory findings were also used to assess the COVID-19 severity for early detection and identification of critically ill patients [16]. Although fatigue was a common COVID-19 symptom, a study in Ireland proved otherwise, as fatigue was shown to have no association between COVID-19 severity (need for inpatient admission, supplemental oxygen or critical care) [17].

Zhang et al. [18] have highlighted those symptoms like muscle ache, shortness of breath and nausea and vomiting, lower lymphocytes, and higher serum creatinine and radiograph score were the predictive factors for the severe or critical subtype. Symptoms like shortness of breath (dyspnoea) were commonly found in most of the COVID-19 severe cases. It is the most significant symptom predictor of COVID-19 death [19]. Neurological manifestations like headache can be presented in COVID-19 as well [20]. Pain related symptoms like myalgia and headaches were found to be high in COVID-19 patients. Furthermore, those with severe cases tend to complain of pain in their bodies. If left untreated, the pain would worsen, leading to chronic pain [21].

Those infected with COVID-19 would also experience gastrointestinal symptoms. Xiao et al. [22] highlighted that those with diarrhoea as their initial symptom also experienced vomiting, nausea, and lack of appetite, respectively. Zhang et al. [23] showed that gastrointestinal symptoms like diarrhoea and lack of appetite in patients were usually accompanied by moderate or high fever. These patients also have a higher probability of developing severe pneumonia. According to Wei et al. [24], COVID-19 patients who had diarrhoea would also experience longer discomfort compared to patients that do not have diarrhoea.

Besides, there were several studies that reported loss of smell and taste conditions. Maechler et al. [25] stated that early-onset presentation of COVID-19 resembled flu-like symptoms except for smell and/or taste dysfunction. Loss of smell was known as olfactory dysfunctions whereas the loss of taste was known as gustatory functions. According to Fantozzi et al. [26], taste alterations were prevalent, followed by xerostomia and smell dysfunctions. Further, COVID-19 patients also experienced anosmia [27]. Sakalli et al. [28] had determined the frequency and severity of general and ear nose throat (ENT)- related symptoms especially smell and/or loss of sense of taste (STL) in COVID-19 disease. This was accompanied by an investigation into the recovery process of STL. Amer et al. [29] highlighted that those with hyposmia recovered faster than those with anosmia. Their findings also evident that females with COVID-10 have a better recovery in regaining back their sense of smell.

Ceron et al. [30] were able to determine the diagnostic accuracy of self-reported smell and taste loss for diagnosing COVID-19. Accordingly, it was easier to determine the diagnostic accuracy of self-reported smell and taste loss among patients who were COVID-19 positive. Xu et al. [31] had compared the clinical characteristics and dynamics of viral load between imported and non-imported patients with COVID-19. Tests were administered to those who are close contacts of COVID-19 cases. Klopfenstein et al. [32] described the prevalence and features of anosmia and dysgeusia in COVID-19 patients. Lechien et al. [33] suggested that dysphonia can be added to the symptom list in COVID-19 as it may exist in patients with severe clinical COVID-19 presentation. Sheng et al. [34] even suggested that dysosmia can be added to the COVID-19 symptom list.

Sudden dysfunction is quite common in patients who were infected with COVID-19. Olfactory dysfunctions can also be found in asymptomatic patients and the olfactory score (OFS) as reported in Bhattacharjee et al. [35]. They found that the area under curve (AUC) value of 0.83 was a useful predictor in asymptomatic patients hence proving that OFS is a reliable method. A study by Liang et al. [36] further proved that neurosensory dysfunction can be an indicator of early diagnosis of COVID-19.

A risk-stratification model was developed by Ryan et al. [38] to predict severe COVID-19 using only the symptoms, comorbidities, and demographic data. The model had the capability of reducing the nosocomial spread and at the same time ensuring high-risk patients receive appropriate care. It was also designed to minimized overfitting by careful selection of candidate variables of high clinical relevance and sufficient event occurrence. There was also a robust SVM-based model developed to predict the opportunity of the patients' progress into severe/critical symptoms using more than 200 clinical and laboratory features [69]. Qiu et al. [37] highlighted the capability of big data and artificial intelligence concepts with the aid of convenient smartphones for disease control. On the other hand, Nomura et al. [70] highlighted that social network services can provide real-time data collection enabling an early warning system for infectious disease. Nevertheless, not many studies had adopted data mining approaches to assess and model the severity of the COVID-19 based on the reported symptoms.

2.5 Attributing Factors to COVID-19 Severity

Past studies have observed various attributing factors to COVID-19. Among the factors considered, age and underlying health conditions were responsible for the severity of the disease (Table 2.2 and Figure 2.3).

Age category: Older patients and individuals with underlying medical health conditions (hypertension, diabetes, cardiovascular disease, pulmonary disease, cerebrovascular disease, dementia, cancer, liver and kidney disease) were at higher risk of developing severe symptoms leading to a higher fatality rate [41], [47], [49], [51]. Wang et al. [53] also showed that old age, chronic basal diseases, and smoking history may be the risk factors for a poor prediction of the course of the disease. Meanwhile, Mei et al. [39]

showed that young patients had more common upper respiratory symptoms and had better recoveries from the disease. These young patients tend to be asymptomatic as well.

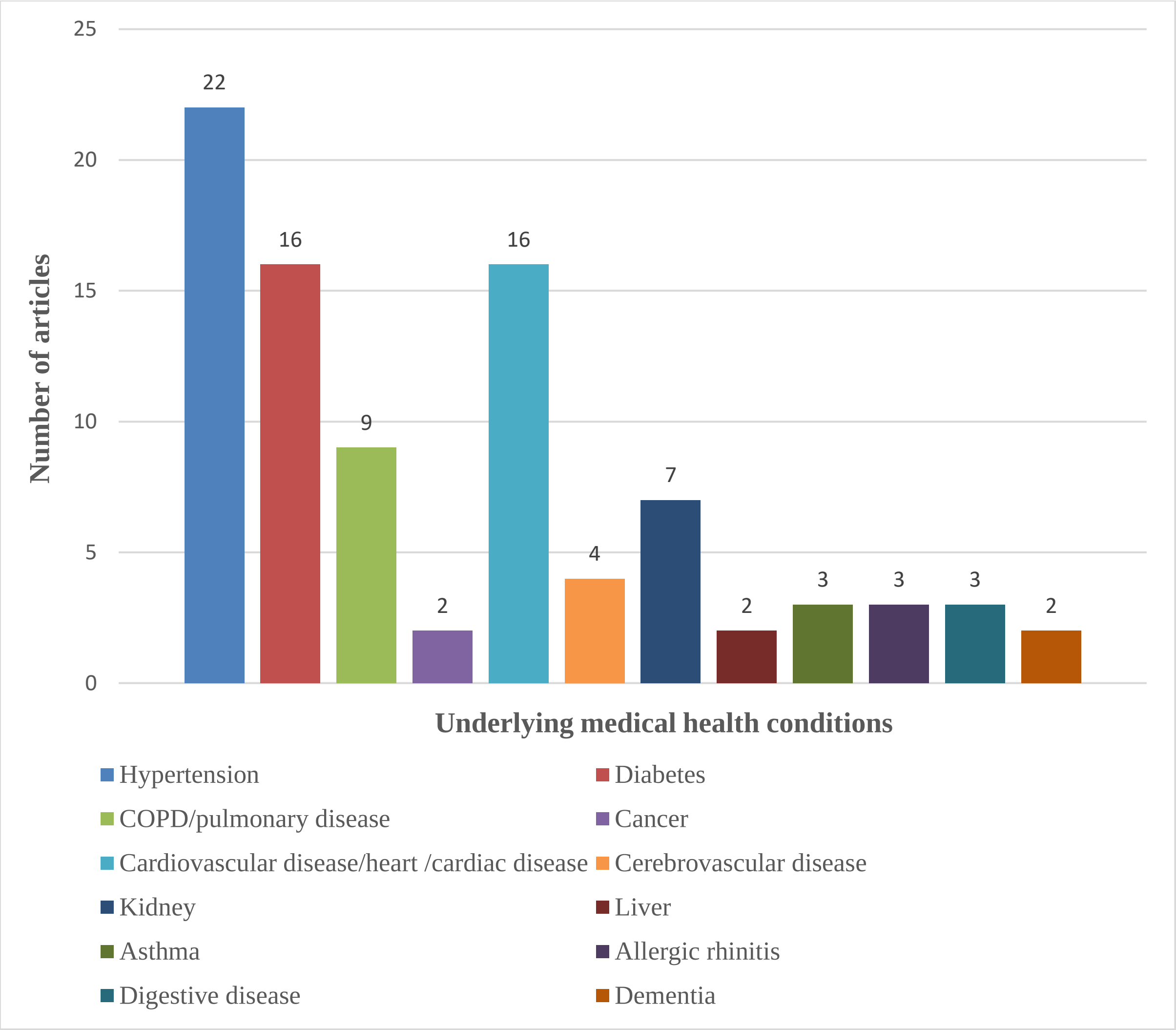


Figure 2.3: Underlying health conditions of patients reported in the literature

Table 2.2: Articles that reported on underlying health conditions associated with COVID-19 disease

Underlying health conditions	**References**
Hypertension	[1], [4], [20], [22], [23], [30]–[32], [34]–[37], [7], [40], [48], [51], [9], [10], [12], [14], [15], [17], [19]
Diabetes	[4], [6], [23], [30]–[32], [34], [36], [51], [10], [12], [14], [15], [17], [19], [20], [22]

Chronic Obstructive Pulmonary Disease (COPD)/ pulmonary disease	[1], [4], [14], [20], [21], [23], [28], [30], [47]
Cardiovascular disease/heart or cardiac disease/coronary disease	[4], [6], [25], [30], [34], [36], [37], [39], [9], [14], [15], [17], [19], [20], [22], [23]
Cerebrovascular disease	[7], [20], [37], [45]
Kidney disease/renal disease	[4], [9], [10], [23], [30], [31], [51]
Digestive disease	[39], [45], [48]
Asthma	[6], [23], [47]
Allergic rhinitis	[3], [12], [48]
Cancer	[9], [23]
Liver disease	[7], [9]
Dementia	[4], [10]

Hypertension: Hypertension is among the most prevalent medical condition in COVID-19 patients (Figure 2.3 and Table 2.2). Hypertension patients were more likely to develop severe pneumonia, excessive inflammatory reactions, organ, and tissue damage as well as the worsening of the disease [57]. Moreover, it is the interaction or the coexistence between SARS-CoV-2 and the angiotensin-converting enzyme (ACE)-2 that leads to the pathogenesis of hypertension.

Diabetes: Patients with diabetes were also more likely to experience poor clinical outcomes. A study by Zhang et al. [58] showed that patients with diabetes had significantly higher leucocyte and neutrophil counts, and higher levels of fasting blood glucose, serum creatinine, urea nitrogen, and creatine kinase isoenzyme MB at admission compared with those without diabetes. High leucocyte and neutrophil counts would indicate that patients with diabetes had a lower immunity, making them more susceptible to viral infection.

Pulmonary disease: Those who have pulmonary disease and were obese have higher risks of developing severe symptoms of COVID-19 [59]. Patients who have pulmonary diseases or even cardiovascular disease were highly likely to experience further health complications which can lead to death. Nevertheless, a contrary study by An et al. [60] found that pulmonary disease such as chronic obstructive pulmonary disease (COPD) was not associated with the mortality in COVID-19. Further study by Rasilla et al. [46] showed that there were suspicious findings of COVID-19 pneumonia in asymptomatic patients with COVID-19. However, Ma et al. [52] showed that the asymptomatic patients in Jinan, China had laboratory indicators and lung lesions on chest CT that were mild. Meng et al. [40] characterized the CT imaging and clinical course of asymptomatic cases with COVID-19 pneumonia. Ding et al. [43] evaluated lung abnormalities on thin-section computed tomographic (CT) scans in patients with COVID-19 and correlate the findings with the duration of symptoms. Yang et al. [55] highlighted that some patients with COVID-19 had CT images showing lung segments with ground-glass opacity, mixed opacity, patchiness, and consolidation. Various CT lung abnormalities detected in COVID-19 patients include bilateral lung involvement, nodular abnormality, patchy abnormality, ground-glass opacity, lung lesion, pleural effusion, consolidation, and crazy paving pattern as summarized in Table 2.3.

Table 2.3: Summary of CT lung abnormalities of COVID-19 patients

CT Lung Abnormalities	**References**
Bilateral lung involvement	[18], [21], [29], [41]
Nodular abnormality	[30], [35]
Patchy abnormality	[9], [17], [35], [43]
Ground-glass opacity	[9], [14], [35], [43], [45], [47]
Lung lesion	[38], [43], [46]
Pleural effusion/ pleural thickening	[9], [30], [43]
Consolidation	[24], [26], [45], [47]
Crazy paving pattern	[26]

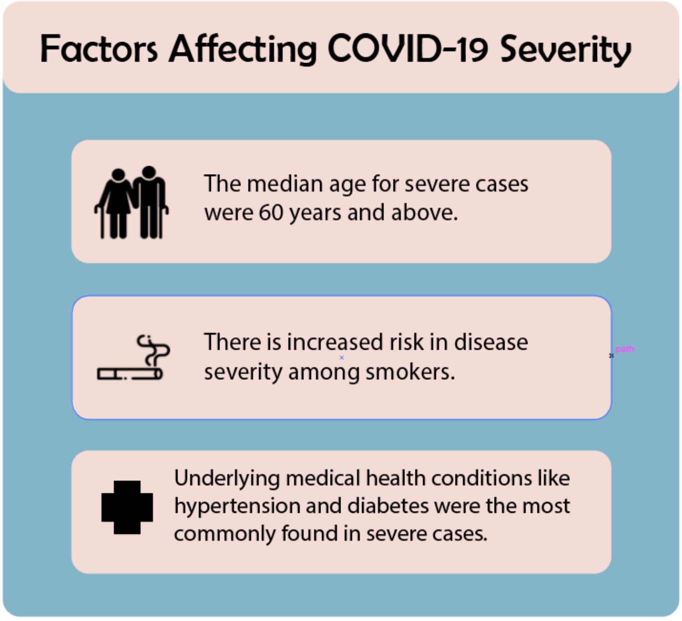
Kidney failure: Apart from these, there are also reports of the prevalence of kidney failure among COVID-19 patients. Patients with pre-existing chronic kidney disease would exhibit lower haemoglobin levels in addition to increased signs of inflammation. This would result in milder respiratory involvement. Furthermore, these groups were treated with anti-COVID-19 medications and invasive ventilation, leading to a higher mortality rate [61]. Other than that, cerebrovascular disease (CVD) was not uncommon among COVID-19 patients. According to Li et al. [62], laboratory findings like elevated C-reactive protein and D-dimer were present in CVD patients compared to those without CVD. This indicates severe inflammation and high coagulation within the patients' body.

Cancer, dementia, and liver disease: Besides that, there are also reports of patients with comorbidities like cancer, dementia, and liver disease. High fatality rates were found in older patients and those with haematological malignancies (leukaemia, lymphoma, myeloma) [63]. In fact, patients of different cancer types differ in their susceptibility to SARS-CoV-2 as well as different COVID-19 phenotypes. Dementia, a neurological disorder commonly found in old patients, was associated with severe COVID-19 and increased mortality [76]**.** Even so, it is still unclear whether the nervous system was harmed via a systemic inflammatory response or virus-induced immunosuppression [64]. A study by Kumar et al. [48] also showed that liver injury is more prevalent in the severe disease group. For patients with chronic liver disease (CLD), the disease progression would be higher. A high level in pro-inflammatory cytokines like IL-6 and ferritin were higher in CLD patients [65].

Allergic rhinitis and asthma: There were minority reports that mentioned allergic rhinitis and asthma (Figure 2.4 and Table 2.2). Allergic rhinitis is the inflammation of the inside of the nose caused by an allergen. Examples of allergens can include pollen, dust, mould, and flakes of skin from animals. This was less frequent due to the use of face masks during the pandemic [67]. Asthma, on the other hand, was not a risk factor for the poor recovery rates in COVID-19 patients. Increasing mortality rates in asthmatic patients are due to acute exacerbation instead [68]. Those who experience acute exacerbation are at higher chances of death even if they are not infected with COVID-19.

Apart from confirmation through the swab tests, other tests to diagnose the virus which include laboratory test like blood tests and CT scan tests were necessary to confirm those who were infected. According to Xia et al. [56], a higher neutrophil to lymphocytes ratio (NLR), lower CD4 counts, and lower RT-PCR-CT values were likely to predict the severity of the COVID-19 disease at the early stage. Besides, Lower RT-PCR-CT values also indicate the infectivity of the patient. These three parameters can be used as the indicator measurements. Furthermore, it is found that the serum level of IL-6 was an effective predictor of disease severity [66].

Overall findings from the review evident the important symptoms were fever, cough, fatigue, dyspnoea, headache, myalgia, diarrhoea, nausea, loss of smell and taste. Less important ones were runny nose and inappetence. Although reports of chest pain were lesser, this symptom sometimes go together with dyspnoea. The more recent works have paid close attention to the super spreaders that are present in asymptomatic and mild cases [45]. In fact, recent variants such as the B.1.1.7 (Alpha), B.1.351(Beta), P.1(Gamma), and B.1.167.2(Delta), spread faster than other variants. Therefore, more attentions should be paid on these circulating variants.



Factors Affecting COVID-19 Severity

The median age for severe cases were 60 years and above.

There is increased risk in disease severity among smokers.

Underlying medical health conditions like hypertension and diabetes were the most commonly found in severe cases.

Figure 2.4: Summary on main factors affecting COVID-19 severity level

2.6 Summary

Initial symptoms presented for COVID-19 were fever and cough. However, fever was later found not incapable to distinguish the disease severity level because high fever temperature does not always equate to severe COVID. Shortness of breath or dyspnoea is a strong indicator of COVID-19 severity as agreed by many studies. Dyspnoea is closely linked to death related COVID-19 cases. Reliance on a particular symptom alone is insufficient to determine if a person is infected with COVID-19. Strong association was realized between risk factors like age and those with underlying health conditions with the COVID-19 severity level. These high-risk groups include the elderly, those with hypertension and diabetes as well as smokers were more susceptible in getting infected by COVID-19. Early detection of the disease is necessary to prevent disease from progressing to a severe state. Thus, to reduce rate of infections, people are urged to adhere to SOPs and to not be complacent towards it.

CHAPTER 3 : METHODOLOGY

3.1 Overview

This chapter focuses on the research methodology that was carried out to by using data mining approach to study the relationship between COVID-19 symptoms and the disease severity in addition to gaining better understanding of this disease. The focus will be on two case studies concerning COVID-19 symptoms retrieved from Kaggle [71] and a research journal from Turkey by Salepci et al. [72]. The entire study involves several stages that include data collection, data pre-processing, model building and data classification as well as verification of the model (Figure 3.1). Data pre-processing is performed to detect and remove potential data outliers and extreme values. Removal of irrelevant attributes will be carried out as well as imputation method on missing values. Several classification algorithms: J48, SMO, Random Forest, and Simple Logistic are executed and tested to classify data into three classes: mild, moderate, and severe. The performance of the algorithms is examined by classification accuracies. A classification predictive model is built to classify COVID-19 severity based on the clinical symptoms.

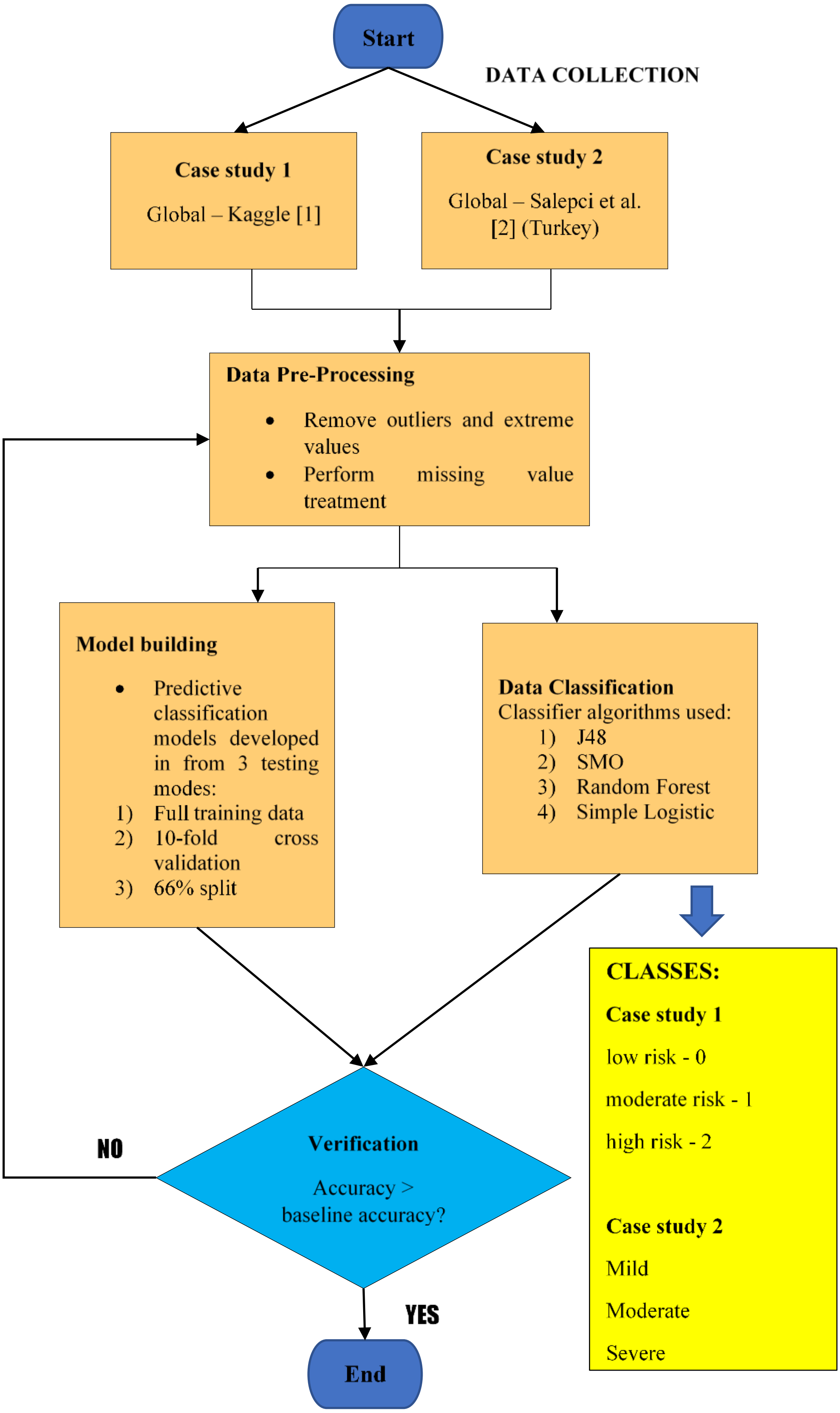


Figure 3.1: Project methodology