**Prediction of COVID-19 using Machine Learning Techniques**

**ABSTRACT**

**Background:** Over the past 4-5 months, the Coronavirus has rapidly spread to all parts of the world. Research is continuing to find a cure for this disease while there is no exact reason for this outbreak. As the number of cases to test for Coronavirus is increasing rapidly day by day, it is impossible to test due to the time and cost factors. Over recent years, machine learning has turned very reliable in the medical field. Using machine learning to predict COVID-19 in patients will reduce the time delay for the results of the medical tests and modulate health workers to give proper medical treatment to them.

**Objectives:** The main goal of this thesis is to develop a machine learning model that could predict whether a patient is suffering from COVID-19. To develop such a model, a literature study alongside an experiment is set to identify a suitable algorithm. To assess the features that impact the prediction model.

**Methods:** A Systematic Literature Review is performed to identify the most suitable algorithms for the prediction model. Then through the findings of the literature study, an experimental model is developed for prediction of COVID-19 and to identify the features that impact the model.

**Results:** A set of algorithms were identified from the Literature study that includes SVM (Support Vector Machines), RF (Random Forests), ANN (Artificial Neural Network), which are suitable for prediction. Performance evaluation is conducted between the chosen algorithms to identify the technique with the highest accuracy. Feature importance values are generated to identify their impact on the prediction.

**Conclusions:** Prediction of COVID-19 by using Machine Learning could help increase the speed of disease identification resulting in reduced mortality rate. Analyzing the results obtained from experiments, Random Forest (RF) was identified to perform better compared to other algorithms.

**CONTENTS**

Abstract i

Acknowledgments iii

1 Introduction 1

1.1 Aim 2

1.2 Objectives ................................. 2

1.3 Research questions ............................ 2

1.4 Defining the scope of the thesis ..................... 2

1.5 Outline ................................... 3

2 Background 5

2.1 Algorithms ................................. 7

3 Related Work

4 Method 11

4.1 Literature Review . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

4.2 Experiment . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

4.2.1 Software Environment . . . . . . . . . . . . . . . . . . . . . . 12

4.2.2 Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13

4.2.3 Data Preprocessing . . . . . . . . . . . . . . . . . . . . . . . . 14

4.2.4 Implementation . . . . . . . . . . . . . . . . . . . . . . . . . . 14

4.2.5 Algorithm Configurations . . . . . . . . . . . . . . . . . . . . 15

4.2.6 Performance Metrics . . . . . . . . . . . . . . . . . . . . . . . 15

References 33

**LIST OF FIGURES**

2.1 Support Vector Machine....................... 7

2.2 Neural Network ............................ 8

2.3 Visualization of Random Forest making a prediction. ....... 8

**Chapter 1**

**Introduction**

Corona viruses are a large family of viruses that are known to cause illness ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome(MERS) and Severe Acute Respiratory Syndrome(SARS) [6]. These two diseases are spread by the corona viruses named as MERS-CoV and SARS-CoV. SARS was first seen in 2002 in China and MERS was first seen in 2012 in Saudi Arabia [8]. The latest virus seen in Wuhan, China is called SARS-COV-2 and it causes corona virus. A pneumonia of unknown cause detected in Wuhan, China was first reported to the World Health Organisation (WHO) Country Office in China on 31 December, 2019 [1]. Since, then the number of cases of corona virus are increasing along with high death toll. Corona virus spread from one city to whole country in just 30 days [50]. On Feb 11, it was named as COVID-19 by World Health Organisation (WHO)[5]. As this COVID-19 is spread from person to person, Artificial intelligence based electronic devices can play a pivotal role in preventing the spread of this virus. As the role of healthcare epidemiologists has expanded, the pervasiveness of electronic health data has expanded too [13]. The increasing availability of electronic health data presents a major opportunity in healthcare for both discoveries and practical applications to improve healthcare [48]. This data can be used for training machine learning algorithms to improve its decision-making in terms of predicting diseases. As of May 16, 2020, totally 44,25,485 cases of COVID-19 have been registered and total number of deaths are 3,02,059 [3]. COVID-19 has spread across the globe with around 213 countries and territories affected [2]. As the rise in number of cases of infected corona virus quickly outnumbered the available medical resources in hospitals, resulted a substantial burden on the health care systems [44]. Due to the limited availability of resources at hospitals and the time delay for the results of the medical tests, it is a typical situation for health workers to give proper medical treatment to the patients. As the number of cases to test for corona virus is increasing rapidly day by day, it is not possible to test due to the time and cost factors [25]. In our thesis, we would like to use machine learning techniques to predict the infection of corona virus in patients.

* 1. **Aim**

The aim of this thesis is to predict whether a person has COVID-19 or not, using machine learning techniques. The prediction is performed using the clinical information of the patients. The goal is to identify whether a patient can potentially be diagnosed with COVID-19.

* 1. **Objectives**

The main objective of our thesis are, • Identifying the most suitable machine learning technique for prediction, to perform on clinical reports of patients. • Preparing a machine learning model that could make accurate predictions of COVID-19 in patients. • Identifying the features that affects the prediction of COVID-19 in patients.

**Chapter 2**

**Background**

Machine Learning is a subset of Artificial Intelligence(AI) and was evolved from pattern recognition where the data can be structured for the understanding of the users. Recently, many applications have been developed using Machine Learning in various fields such as healthcare, banking, military equipment, space etc. Currently, Machine Learning is a rapidly evolving and continuously developing field. It programs computers using data to optimize their performance. It learns the parameters to optimize the computer programs using the training data or its past experiences. Using the data, it can also predict the future. Machine Learning also helps us in building a mathematical model using the statistics of the data. The main objective of Machine Learning is that it learns from the feed data without any interference of humans that is, it automatically learns from given data(experience) and gives us the desired output where it searches the trends/patterns in the data[43].

It is broadly classified into four types:

• Supervised Machine Learning.

• Unsupervised Machine Learning.

• Semi-Supervised Machine Learning.

• Reinforcement Machine Learning.

**2.1 Supervised Machine Learning:**

Supervised Learning is a Machine Learning model that is built to give out predictions. This algorithm is performed by taking a labelled set of data as input and also known responses as output to learn the regression/classification model. It develops predictive models from classification algorithms and regression techniques.

Classification predicts discrete responses. Here, the algorithm labels by choosing two or more classes for each example. If it is done between two classes then it is called binary classification and if it is done between two or more classes then it is called multi- class classification. Applications of classification includes hand writing recognition, medical imaging etc.

Regression predicts continuous responses. Here, the algorithms returns a statistical value. For example, a set of data is collected such that the people are happy when considered the amount of sleep. Here, sleep and happy are both variables. Now, the analysis is done by making predictions[11].

The types of popular regression techniques are:

• Linear regression. • Logical regression.

**2.2 Unsupervised Machine Learning:**

Unlike the supervised learning, there is no supervisor here and we only have input data. Here, the basic aim is to find certain patterns in the data that occur more than others. According to the statistics, it is called density estimation. One of the methods for the density estimation is called clustering. Here, the input data is formed into clusters or groupings. Here, the assumptions are made such that the clusters are discovered which will match reasonably well with a classification. This is a datadriven approach that works better when provided with sufficient data. For example, the movies in Netflix.com are suggested based on the principal of clustering of movies where several similar movies are grouped based on customer’s recently watched movie list. It mostly discovers the unknown patterns in the data but most of the time these approximations are weak when compared with the supervised learning[12].

**2.3 Semi-supervised Machine Learning:**

The name “semi-supervised learning” comes from the fact that the data used is between supervised and unsupervised learning [57]. Semi-supervised algorithm has the tendency to learn both from labelled and unlabelled data. Semi-supervised machine learning gives high accuracy with a minimum annotation work. Semi-supervised machine learning uses mostly unlabelled data together combined with labelled data to give better classifiers. As less annotation work is enough to give good accuracy, humans have less work to do here.

**2.4 Reinforcement Machine Learning:**

Reinforcement learning learns its behaviour from a trial and error method in a dynamic environment. Here, the problem is solved by taking an appropriate action in a certain situation to maximize the output and to obtain the acquired results. In Reinforcement Learning, there is presentation of the input or output data. Instead, when the desired action is chosen, the agent is immediately told the reward and the next state are not considering the long terms actions. For the agent to act optimally it should have the knowledge about states, rewards, transitions and actions actively.

**Chapter 3**

**Related Work**

Nanshan Chen et al. performed a retrospective, single-centre study of various patients data from Jinyintan Hospital in Wuhan, China. In this research they described the epidemiological data(short term) or long term exposure to virus epicenters, signs and symptoms, laboratory results, CT Findings and clinical outcomes[16]. Though this research does not directly focus on the prediction of COVID-19, it gives us a better understanding of the clinical outcomes.

Shuai Wang et al. has identified the radio-graphical changes in CT images of patients suffering from COVID-19 in China. In this research, he has used deep learning methods to extract COVID-19’s graphic features through the CT scan images to develop it as a alternative diagnostic method. They have collected CT images of confirmed COVID-19 Patients along with those who were diagnosed with pneumonia. The results from their work provide the proof-of-principle for the use of AI for accurate COVID-19 prediction[47]. This research uses CT Scan images, which is different from our research as we use clinical features and laboratory results for the prediction.

Dawei Wang et al. in this research has described the epidemiological, demographic, clinical, laboratory, radio-logical and treatment data from Zhongnan Hospital, Wuhan China. The data was analysed and documented to be used to track the infections[46]. The author gives better insights about the radio-logical and treatment data that could be used for our prediction of COVID-19 in our model.

Halgurd S. Maghdid et al. have proposed a new framework to detect corona virus disease using the inboard smartphone sensors. The designed AI framework collects data from various sensors to predict the grade of pneumonia as well as predicting the infection of the disease [26]. The proposed framework takes uploaded CT Scan images as the key method to predict COVID-19. This framework relies on multireadings from multiple sensors related to the symptoms of COVID-19.

Ali Narin et al. has developed an automatic detection system as an alternative diagnosis option of COVID-19. In this study, "three different convolutional neural network based models (ResNet50, InceptionV3 and Inception-ResNetV2) have been proposed for the detection of corona virus pneumonia infected patient using chest X-ray radio graphs [32]". The author also discusses about the classification performance accuracy between the three CNN models.

In [52], the authors proposed a three-indices based model to predict the mortality risk. They built a prognostic prediction model based on XGBoost machine learning algorithm to predict the mortality risk in patients. They determined a clinical route which is simple to check and asses the risk of death. The research focuses on the mortality risk which is different from our research, where the prediction is completely based on the clinical findings of patients suffering from COVID-19.

The authors in the article [9], presented a comparative analysis of machine learning models to predict the outbreak of COVID-19 in various countries. Their study and analysis demonstrate the potential of machine learning models for the prediction of COVID-19. The article was based entirely on the outbreak of cases in various countries. In our work we predict the disease by using the clinical information.

In [38], the authors performed bench-marking evaluation of various machine learning algorithms, deep learning algorithms and various ICU scoring systems on various clinical prediction tasks. This task was conducted with publicly available clinical data sets. In our work we specifically work on the COVID-19 patient information.

In the above mentioned papers, various prediction systems were developed using CT Scan images and symptoms for prediction of COVID-19, mortality risks, outbreak in various countries. As per the existing knowledge, there is not much evidence of prediction system using clinical information. This thesis will be using machine learning techniques to predict COVID-19 with clinical information of patients suffering from COVID-19. It will also determine which features would impact the prediction model.

**Chapter 4**

**METHODOLOGY**

The research methods we used here are Literature review and Experiment. Firstly, we performed a systematic literature review where we carefully analysed the literature and from the results we conducted an experiment for research question 1 through which we identified suitable machine leaning techniques for prediction. For research question 2, we conducted an experiment, where we determined what features would influence the results of the prediction of COVID-19.

**4.1 Experiment**

An experiment is conducted with the results achieved from the SLR (Systematic Literature Review) to reach the goals of RQ1 where we identify the suitable machine learning technique for prediction of COVID-19. The experiment is further continued to build a model of prediction with the selected algorithm to determine RQ2 where the factors that influence the prediction are identified.

**4.1.1 Software Environment**

Python Python is a high level and effective general use programming language. It supports multi-paradigms. Python has a large standard library which provide tools suited to perform various tasks. Python is a simple, less-clustered language with extensive features and libraries.

Different programming abilities are utilized for performing the experiment in our work. In this thesis, the following python libraries were used [45].

• Pandas - It is a python package that provides expressive data structures designed to work with both relational and labelled data. It is an open source python library that allows reading and writing data between data structures [30].

• Numpy - It is an open source python package for scientific computing. Numpy also adds fast array processing capacities to python [29].

• Matplotlib - It is an open source python package used for making plots and 2D representations. It integrates with python to give effective and interactive plots for visualization [29].

• Tensorflow - It is a mathematical open source python library designed by Google Brain Team for Machine intelligence [55].

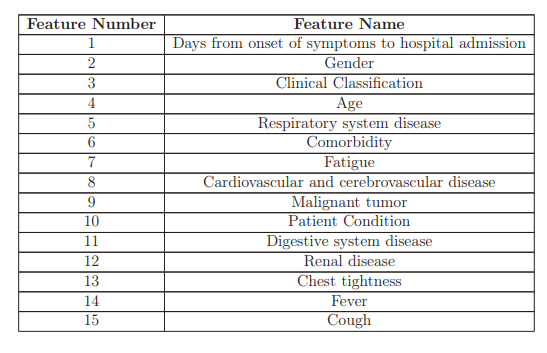
• Sklearn - It is an open source python machine learning library designed to work alongside Numpy. It features various machine learning algorithms for classification, clustering and regression.

**4.1.2 Dataset Data Collection:**

Data collection was an essential and protracted process. Regardless the field of research, accuracy of the data collection is essential to maintain cohesion. As the clinical information of patients was not publicly available, it was an inflexible and tedious process to collect the data. Various Hospitals and Health Institutes in Sweden and China were approached to get the most accurate data but due to the present situation at hospitals with heavy inflow of patients with COVID-19, we couldn’t get access to direct information. An intense search was conducted on various databases to gather open source clinical information of patients diagnosed with COVID-19.

Dataset Used The data set that was used to train the model to predict COVID-19 was gathered from an open source data shared by Yanyan Xu at a repository figshare[51]. The data set contained information about hospitalized patients with COVID-19. It included demographic data, signs and symptoms, previous medical records, laboratory values that were extracted from electronic records. To train the model with equal records of patients with negative samples another data set from Kaggel repository was used[4]. The original data-set contained details of medications followed by the doctors to cure the disease. As our model doesn’t require such data, those fields have been eliminated. The data-set is a combined multi-dimensional data. Some of the data gives information whether the patient is diagnosed with a particular disease in the past such as Renal Diseases, Digestive Diseases and other data contains precise clinical values obtained previously. It contains fields with textual data and some with precise values. Textual data was encoded with integer values for experimental setup. The attributes that were considered in the data-set for the machine learning model are presented in Table 4.1.

Table 4.1: Features in the dataset used.



**4.1.3 Data Preprocessing:** Data preprocessing is an important process in development of machine learning model. The data collected is often loosely controlled with out-of-range values, missing values, etc. Such data can mislead the result of the experiment.

• Imputation of missing values - In our data, missing values have been handled by using simple imputer from sklearn python package. The missing values are replaced by using mean strategy.

• Encoding Categorical Data - We used the package of OneHotEncoder in python, this package handles categorical data by one-hot or dummy encoding scheme.

**4.1.4 Implementation:** The experiment was conducted in the Python IDLE, which is a default integrated development and learning environment for python.

The experiment was conducted in various phases that are mentioned below:

• After data collection, the patients data is divided into record sets containing 100 records, 150 records, 200 records, 250 records, 300 records, 355 records respectively.

• A 5-fold cross validation technique is used to randomize the testing data-set to get accurate results. Experiment on each machine learning algorithm is conducted by 5-fold cross validation with each of the record sets.

• The prediction accuracy of each algorithm at each record set is compared and evaluated for selecting the suitable algorithm for this data-set.

• A feature importance experiment is conducted to evaluate the importance of each attribute on the artificial classification task.

**4.1.5 Performance Metrics:** It is an essential task to measure the performance of a machine learning model. As our model requires classification, we have used accuracy as the performance metric. Accuracy Accuracy is the metric used in this thesis for evaluation of the algorithms. It is the most used performance metric to evaluate classification techniques. This measure allows us to understand which model is best at identifying patterns in training set to give better predictions in the unknown test data-set.

Accuracy = T P + T N/ T P + T N + F P + F N

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

**REFERENCES**

[1] Coronavirus Disease (COVID-19) - events as they happen. Library Catalog: [www.who.int](http://www.who.int).

[2] Countries where Coronavirus has spread - Worldometer. Library Catalog: [www.worldometers.info](http://www.worldometers.info).

[3] COVID-19 situation reports. Library Catalog: [www.who.int](http://www.who.int).

[4] Diagnosis of covid-19 and its clinical spectrum dataset. url=https://kaggle.com/einsteindata4u/covid19.

[5] WHO Director-General’s remarks at the media briefing on 2019-nCoV on 11 February 2020. Library Catalog: [www.who.int](http://www.who.int).

[6] WHO EMRO | Questions and answers | COVID-19 | Health topics.

[7] Support Vector Machine Machine learning algorithm with example and code, January 2019. Library Catalog: www.codershood.info Section: Machine learning.

[8] Ali Al-Hazmi. Challenges presented by MERS corona virus, and SARS corona virus to global health. Saudi journal of biological sciences, 23(4):507–511, 2016. Publisher: Elsevier.

[9] Sina F Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M Atkinson. Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188, 2020.

[10] Hiba Asri, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science, 83:1064–1069, 2016.

[11] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. New advances in machine learning, pages 19–48, 2010.

[12] Taiwo Oladipupo Ayodele. Types of machine learning algorithms. New advances in machine learning, pages 19–48, 2010. Publisher: InTech.

[13] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage highrisk and high-cost patients. Health Affairs, 33(7):1123–1131, 2014. 33 34 .

[14] Hetal Bhavsar and Amit Ganatra. A comparative study of training algorithms for supervised machine learning. International Journal of Soft Computing and Engineering (IJSCE), 2(4):2231–2307, 2012.

[15] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning, pages 161–168, 2006.

[16] Nanshan Chen, Min Zhou, Xuan Dong, Jieming Qu, Fengyun Gong, Yang Han, Yang Qiu, Jingli Wang, Ying Liu, Yuan Wei, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in wuhan, china: a descriptive study. The Lancet, 395(10223):507–513, 2020. [17] Dursun Delen. Analysis of cancer data: a data mining approach. Expert Systems, 26(1):100–112, 2009.

[18] Manoj Durairaj and Veera Ranjani. Data mining applications in healthcare sector: a study. International journal of scientific & technology research, 2(10):29– 35, 2013.

[19] Ashok Kumar Dwivedi. Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Computing and Applications, 29(10):685–693, 2018.

[20] Arihito Endo, Takeo Shibata, and Hiroshi Tanaka. Comparison of seven algorithms to predict breast cancer survival (< special issue> contribution to 21 century intelligent technologies and bioinformatics). International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association, 13(2):11–16, 2008.

[21] Ya-Han Hu, Yi-Lien Lee, Ming-Feng Kang, and Pei-Ju Lee. Constructing inpatient pressure injury prediction models using machine learning techniques. Computers, Informatics, Nursing: CIN, 2020.

[22] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. Journal of artificial intelligence research, 4:237–285, 1996.

[23] G. Kesavaraj and S. Sukumaran. A study on classification techniques in data mining. In 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), pages 1–7, 2013.

[24] Barbara Kitchenham. Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004):1–26, 2004.

[25] Halgurd S. Maghdid, Kayhan Zrar Ghafoor, Ali Safaa Sadiq, Kevin Curran, and Khaled Rabie. A novel ai-enabled framework to diagnose coronavirus covid 19 using smartphone embedded sensors: Design study. arXiv preprint arXiv:2003.07434, 2020.