Bachelor of Science in Computer Science
May 2020

Prediction of COVID-19 using Machine Learning Techniques

Durga Mahesh Matta
Meet Kumar Saraf

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfilment of the requirements for the degree of Bachelor of Science in Computer Science. The thesis is equivalent to 10 weeks of full time studies.

The authors declare that they are the sole authors of this thesis and that they have not used any sources other than those listed in the bibliography and identified as references. They further declare that they have not submitted this thesis at any other institution to obtain a degree.

Contact Information:
Author(s):
Durga Mahesh Matta
E-mail: duma19@student.bth.se

Meet Kumar Saraf
E-mail: mesa19@student.bth.se

University advisor:
Suejb Memeti
Department of Computer Science

Faculty of Computing
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

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.

Keywords: COVID-19, Machine Learning, Prediction, Supervised Learning, Classification Techniques

Acknowledgments

We would like to show our sincere gratitude to Prof. Suejb Memeti for supervising our thesis and guiding us throughout the project with quick and helpful feedback.

We would also like to thank our dear friend Akhila Dindi for providing constructive comments, which helped in improving our work and would like to extend our gratitude to all those who helped us directly and indirectly.

Contents

List of Figures

List of Tables

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.

1.1 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.

1.2 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.

1.3 Research questions

To achieve the objectives of our thesis, there are some research questions that have been formulated:

1. Which suitable machine learning technique can be used to predict COVID-19?

Motivation: The motivation of the research question is to conduct a conjunctive literature study and experiment to see what are the appropriate machine learning algorithms that can be best applied to the given data and also to find out which algorithm gives us the best results in predicting COVID-19.

2. What are the features that will influence the predictive result of COVID-19?

Motivation: The motivation of this research is to conduct an experiment to identify the features that will influence the results of prediction of Corona virus in human beings.

1.4 Defining the scope of the thesis

This research focuses on development of a machine learning model for predicting COVID-19 in patients. We also work to identify the features from the clinical information of patients that would influence the predictive result of COVID-19. This study does not focus on outer factors such as weather or any environmental factors that might influence results.

1.5 Outline

The thesis structure is divided into different chapters which are as follows:

- Chapter 1: This chapter contains the introduction to this thesis, aim and objectives, research questions, and motivation.

- Chapter 2: In this chapter, we discuss the background of the concepts used during the research.

- Chapter 3: This chapter contains the summary of the works similar to this thesis.

- Chapter 4: This contains methods to answer research questions. It includes experimental analysis like data processing, tools used during the experiment, and experimental setup details.

- Chapter 5: Results obtained are presented in this chapter.

- Chapter 6: This chapter consists of analysis and discussions about the results and methods, the contribution of the thesis to the existing research, threats to the validity of the thesis.

- Chapter 7: In this chapter,we discuss the conclusion of the thesis and discussion on possible future work.

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.

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.~~

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 data-driven 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].

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.

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. Formally,the model consists of [22]:

- a discrete set of environment states, S;
- a discrete set of agent actions, A;
- a set of scalar reinforcement signals; typically {0;1} or the real numbers.

2.1 Algorithms

During our research, we have investigated three algorithms through which we have performed supervised classification.

Support Vector Machines(SVM)

Support Vector Machines performs classification by constructing N-dimensional hyper plane that separates the data into two categories [12]. In SVM, the predictor variable is called an attribute and the transformed attribute is called a feature. ~~Selecting the most suitable representative data is called feature selection.~~ A set of feature describing one case is called a vector.

The ultimate goal of SVM modelling is to find the optimal hyper plane that separates the clusters where on one side of the plane there is target variable and on the other side of the plane other category. The vectors which are near the hyper plane are the support vectors[12].In Figure 2.1, a typical example of support vector machine is depicted.

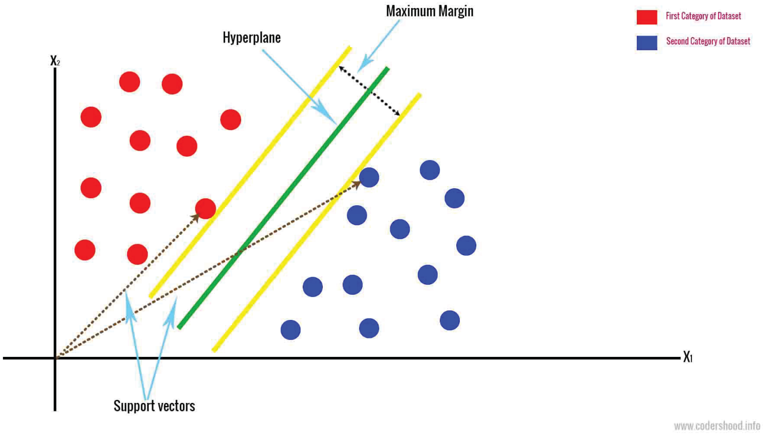


Figure 2.1: Support Vector Machine[7]

Artificial Neural Networks(ANN)

ANNs are an attempt, in the simplest way, to imitate the neural system of the human brain [53]. The basic unit of ANN are neurons. A neuron is said to perform functions on an input and produces an output [56]. Neurons combined together are called neural networks. Once the neural networks are formed, training of the data is started to minimize the error. In the end, an optimizing algorithm is used to further reduce the errors. The layered architecture of Artificial Neural Networks (ANNs) is represented in Figure 2.2.

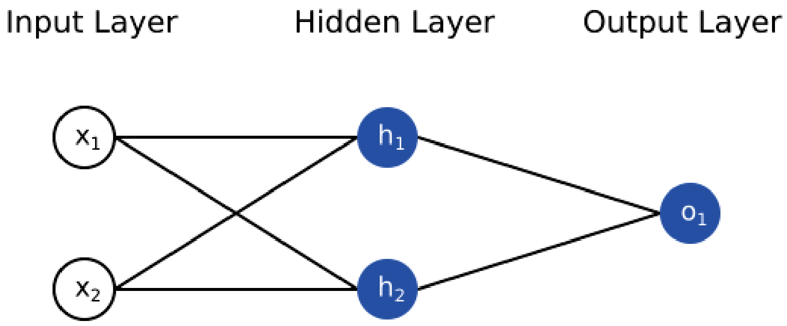


Figure 2.2: Neural Network [56]

Random Forests(RF)

The random sampling and ensemble strategies utilized in RF enable it to achieve accurate predictions as well as better generalizations [40]. The random forests consists of large number of trees. The higher the number of uncorrelated trees, the higher the accuracy [54]. Random Forest classifiers can help filling some missing values. Prediction in Random Forests (RFs) is represented in Figure 2.3.

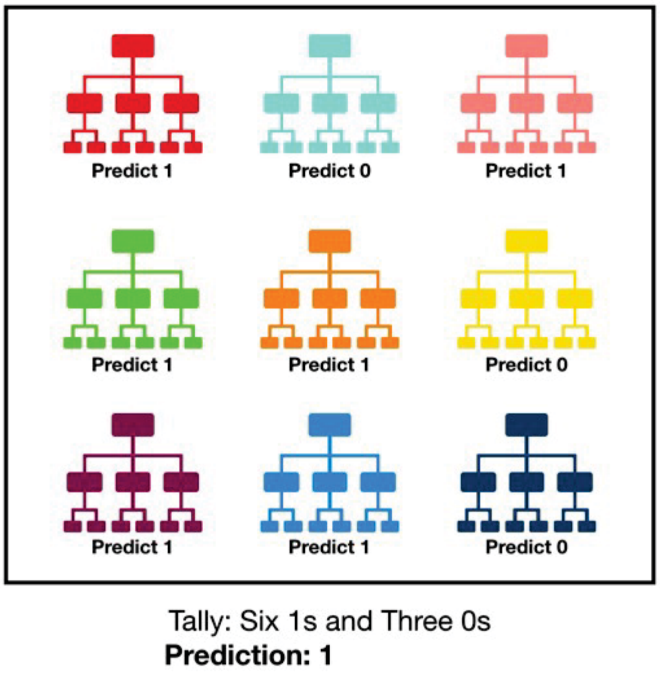


Figure 2.3: Visualization of Random Forest making a prediction. [54]

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 multi-readings 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

Method

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 Literature Review

A systematic literature through the guidelines of Claes Wohlin[49] and Barbara Kitchenham[24], has been conducted to analyze and answer RQ1. This literature review focuses on the understanding of several machine learning algorithms and also identifying appropriate machine learning algorithms that can be used for prediction. There are several steps that we performed in our research, which are:

1. **Identifying the key words**: We have identified the following keywords which are Supervised Machine Learning algorithms, COVID19, classification, prediction.

2. **Formulating the search strings**: From the above identified keywords, primary keywords were selected to formulate the search string.

3. **Locating the literature**: Using search string, the search was performed on various digital database platforms such as Google scholar, IEEE and Science Direct.

4. **Following the Inclusion and Exclusion criteria for selection**: From the collected literature such as articles and conference papers, the inclusion and exclusion criteria is implemented to confine our research.

   Inclusion Criteria

   - Papers related to prediction of COVID-19 using Machine Learning algorithms.
   - All articles should be in English language.

   Exclusion Criteria

   - Incomplete articles.

  - Articles not in English are not considered.

5. **Evaluating and selecting the literature**: After the implementation of the inclusion and exclusion criteria, further the refining is done through careful evaluation and selection of the gathered literature.

6. **Summarizing the literature**: The overall findings from the gathered literature is summarized and represented for analysis.

4.2 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.

4.2.1 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.

4.2.2 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.

Feature Number	Feature Name
1	Days from onset of symptoms to hospital admission
2	Gender
3	Clinical Classification
4	Age
5	Respiratory system disease
6	Comorbidity
7	Fatigue
8	Cardiovascular and cerebrovascular disease
9	Malignant tumor
10	Patient Condition
11	Digestive system disease
12	Renal disease
13	Chest tightness
14	Fever
15	Cough

16	Liver disease
17	Endocrine system disease
18	Diarrhea Chest
19	CT findings - Advances, Absorption
20	White Blood Cell Count
21	Neutrophil count
22	Lymphocyte count
23	Monocyte count
24	CRP - C-reactive protein
25	PCT - Procalcitonin

Table 4.1: Features in the dataset used.

4.2.3 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.

4.2.4 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.

4.2.5 Algorithm Configurations

In this section, the configuration of the algorithms is mentioned. Changes made to the configuration of the algorithm can effect the results.

- Support Vector Machines:

  SVC(kernel = ’linear’, random_state = 0)

- Artificial Neural Networks:

  Layers:
  ann.add(tf.keras.layers.Dense(units=6, activation=’relu’))
  ann.add(tf.keras.layers.Dense(units=6, activation=’relu’))
  ann.add(tf.keras.layers.Dense(units=1, activation=’sigmoid’))

  Compiling the ANN:
  ann.compile(optimizer = ’adam’, loss = ’binary_crossentropy’, metrics = [’accuracy’])

- Random Forests:

  RandomForestClassifier(n_estimators = 10, criterion = ’entropy’, random_state = 0)

4.2.6 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.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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

Chapter 5

Results

5.1 Literature Review Results

A Systematic Literature Review (SLR) is conducted in order to answer RQ1. Which machine learning technique can be used to predict COVID-19? The goal of the SLR is to identify the most suitable algorithms that would facilitate for accurate prediction of COVID-19.

Title	Findings
Supervised machine learning algorithms: classification and comparison [34].	This paper determines the most efficient classification algorithm based on a clinical data-set (Diabetics). Seven supervised machine learning algorithms were considered concluding SVM (Support Vector Machines) followed by RF (Random Forests) that were found with most precision and accuracy [34].
Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in E-health [27].	The author stated that no single supervised algorithm can outperform other algorithms over all data-sets. The simplest approach is to estimate the accuracy of the algorithms and choose the suitable one. But in general, SVM (Support Vector Machines) and ANN (Artificial Neural Networks) tend to perform better when dealing with multi-dimensional and continuous features [27].
An empirical comparison of supervised learning algorithms [15].	Of all the six algorithms that were compared in this paper, Calibrated Boosted trees, Random Forests give best performance in all metrics. Artificial Neural Networks has reached its peak performance with large datasets [15].
Performance evaluation of different machine learning techniques for prediction of heart disease [19].	Logistic regression acquires highest accuracy among the compared algorithms followed by Artificial Neural Networks. SVM (Support Vector Machines) on the other hand acquires highest precision [19].

Bench marking deep learning models on large healthcare data-sets [38].	In this paper, an exhaustive bench marking evaluation has been performed to demonstrate that deep learning algorithms outperform other approaches when large number of clinical time series data is used for prediction tasks [38].
A comparative study of training algorithms for supervised machine learning [14].	A comparative study classification algorithms like Decision Tree Induction, Bayesian Network, Neural Network, K-nearest neighbours and Support Vector Machine has been conducted to justify that each algorithm has its own field of excellence. They suggested that one can use an algorithm for their data by comparing the metrics they require [14].
Using machine learning algorithms for breast cancer risk prediction and diagnosis [10].	In this paper, a comparison between Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) data-sets is conducted in terms of accuracy and precision. SVM is determined to get the highest accuracy among all other algorithms [10].
Automatic short-term solar flare prediction using machine learning and sunspot associations [39].	Though this paper belongs to a different domain, as accuracy comparison between algorithms is performed it has been considered. Machine learning algorithms such as Cascade-Correlation Neural Networks (CCNNs), Support Vector Machines (SVMs) and Radial Basis Function Networks (RBFN) to conclude that SVM gives highest accuracy. Hybrid model is suggested to be used based on the datasets [39].
Intelligent heart disease prediction system using data mining techniques [35].	In this research a intelligent heart disease prediction system is developed with Decision Trees, Naive Bayes and Artificial Neural Network to compare the performance. The results of the research state that each technique has its own strength in uniquely defined mining goals [35].
Analysis of cancer data: a data mining approach [17].	In this research Decision trees, Artificial Neural Networks, Support Vector Machines and Logistic Regression are compared to develop prediction models for prostate cancer survivability. Support Vector Machines have been found as the most accurate followed by Artificial Neural Networks [17].

Medical data mining and predictive model for colon cancer survivability [33].	A predictive model to predict mortality rate has been designed with Decision Tree, Bayes Networks, and Artificial Neural Network. After the experiment, results show that Artificial Neural Networks give accurate classifications [33].
Analysis of Machine Learning Algorithms on Cancer Dataset [37].	A comparative experiment on Random Forest, Support Vector Machine, Naive Bayes, Decision Tree, Neural Networks and Logistic Regression has been conducted using Weka (Waikato Environment for Knowledge Analysis) tool with Cancer dataset. The results conclude that Support Vector Machines (SVMs) have the highest accuracy followed by Artificial Neural Networks and Random Forest [37].
Analytical Comparison of Machine Learning Techniques for Liver Dataset [41].	A comparative experiment has been conducted on 4 machine learning algorithms trained with liver dataset. The results show that Random Forests is the most suitable algorithm among the others [41].
Predicting the severity of breast masses with data mining methods [31].	The article is a comparative study of Decision Tree (DT), Artificial Neural Network (ANN), and Support Vector Machine (SVM) which are analyzed on mammographic masses dataset. The results summarize that SVM perform with the highest accuracy followed by ANN [31].
Data mining applications in healthcare sector: a study [18].	This article gives a comparison of various data mining techniques, summarizes that no single algorithm can be decided as the most suitable for healthcare sector. They suggested that a comparative experiment must be conducted to get accurate results [18].
A Machine Learning Approach for Early Prediction of Breast Cancer [28].	In this research a comparison experiment on Naive Bayes, Logistic Regression and Random Forest has been conducted using Breast Cancer dataset. The results summarize that Random Forest gives the most accurate predictions [28].
Comparison of seven algorithms to predict breast Cancer survival contribution to 21 century intelligent technologies and bioinformatics [20].	Seven algorithms that include Logistic Regression model, Artificial Neural Network (ANN), Naive Bayes, Bayes Net, Decision Trees with naive Bayes, Decision Trees (ID3) and Decision Trees (J48) have been compared in various metrics. It is stated that Logistic regression model gives highest accuracy followed by Artificial Neural Networks which also has highest precision [20].

A study on classification techniques in data mining [23].	In this article, the experimental results state that it is difficult to choose one algorithm superior to another. It summarizes that classification algorithms are strictly confined to their problem domain [23].
A critical study of selected classification algorithms for liver disease diagnosis [42].	A study of various classification algorithms has been performed through which Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) are summarized as the algorithms with most accuracy and precision [42].
Comparison of machine learning algorithms to predict psychological wellness indices for ubiquitous healthcare system design [36].	A comparison of four machine learning algorithms has been performed and Support Vector Machine (SVM) and K-Nearest Neighbour (KNN) are identified as the best performers [36].
Constructing Inpatient Pressure Injury Prediction Models Using Machine Learning Techniques [21].	A comparative experimental model between Decision Tree, Logistic Regression, and Random Forest has been conducted and identified that Random Forests give the most accurate predictions [21].

Table 5.1: Literature Review Results.

Numerous works using machine learning algorithms in healthcare domain were identified in the Systematic Literature Review (SLR). Most of the articles included a comparison between machine learning techniques. As per [14], [35], [18], [23], healthcare datasets require comparison between algorithms to identify the most suitable one. Support Vector Machines(SVMs), Artificial Neural Networks (ANNs) and Random Forests(RFs) are identified as the most utilized algorithms when accuracy is the performance metric.

5.2 Experiment Results

This chapter presents the results that are obtained from the experiment. The performance metric mentioned in Section 4.2.5 is utilized to evaluate the performance of the algorithms that were selected after the Literature Review. Three algorithms that were identified as the most suitable for the classification task to predict COVID-19 are:

- SVM (Support Vector Machine).
- RF (Random Forests).
- ANN (Artificial Neural Networks).

Each of the above stated algorithms were trained with the data-set that was collected and results were interpreted. Performance of each algorithm was evaluated at different stages of training set. Each algorithm was trained with records sets containing 100 records, 150 records ,200 records, 250 records, 300 records, 355 records respectively. This experiment is performed to obtain which algorithm would be the most suitable for prediction of COVID-19. Also, as the data is split into smaller sets, we could also asses which algorithm would perform better with different datasets available.

5.2.1 Support Vector Machine (SVM) Results

Support Vector Machine (SVM) algorithm is trained with each record sets to identify its accuracy at all stages. At all stages, the data was divided into training and test data by using k-fold cross validation (5-folds). SVM achieves an accuracy of 98.33%. Table 5.2 represent the accuracy for every set of records achieved by Support Vector Machine (SVM) algorithm.

Number of Patient Records	**Accuracy**
100	94.73%
150	96%
200	97.36%
250	97.18%
300	97.71%
355	98.33%

Table 5.2: Support Vector Machine (SVM) Accuracy Results

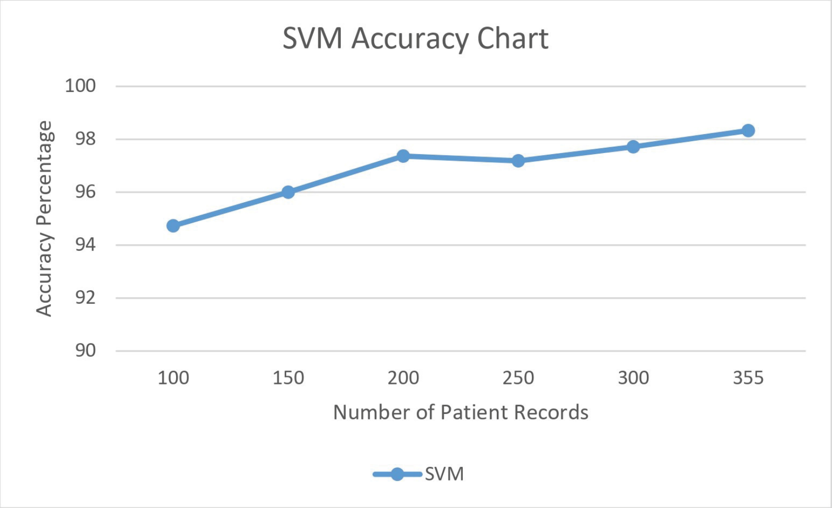


Figure 5.1: Support Vector Machine (SVM) Accuracy Chart

The classification accuracy of Support Vector Machine (SVM) at each record set can be clearly identified from the chart in Figure 5.1

5.2.2 Random Forest (RF) Results

Random Forest (RF) algorithm is trained in a similar way with each records set to identify its accuracy at all stages. At all stages, the data was divided into training and test data by using k-fold cross validation (5-folds). RF achieves an accuracy of 99.44%. The classification accuracy of Random Forest (RF) algorithm for every set of records is represented in Table 5.3.

Number of Patient Records	Accuracy
100	93.33%
150	96.15%
200	96.29%
250	98.36%
300	98.66%
355	99.44%

Table 5.3: Random Forest (RF) Accuracy Results

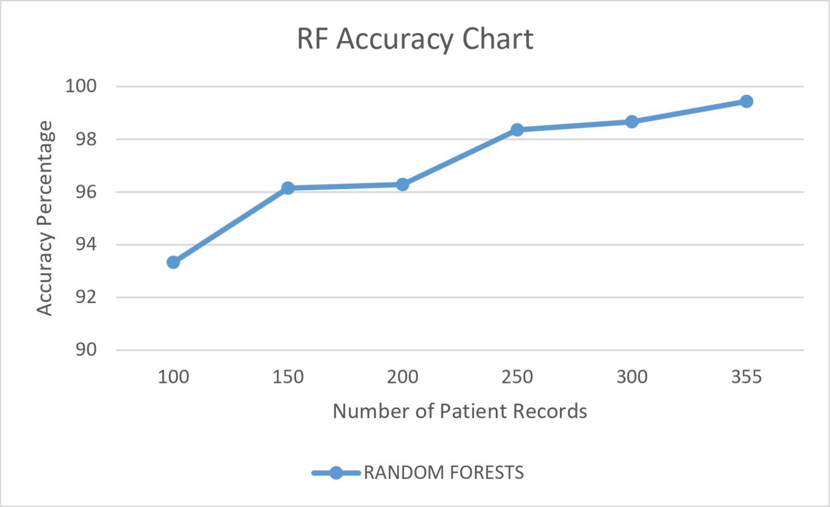


Figure 5.2: Random Forest (RF) Accuracy Chart

The classification accuracy of Random Forest (RF) at each record set can be identified from the chart in Figure 5.2. The figure represents the change in accuracy while using each record set as training data.

5.2.3 Artificial Neural Networks (ANN) Results

Artificial Neural Networks (ANN) Algorithm is trained on data with record sets and tested. On implementing ANN Algorithm, it achieves an classification accuracy of 99.25%. The classification accuracy reported with each record set is tabulated in Table 5.4.

Number of Patient Records	**Accuracy**
100	80.00%
150	86.20%
200	90.90%
250	96.07%
300	98.65%
355	99.25%

Table 5.4: Artificial Neural Networks (ANN) Accuracy Results

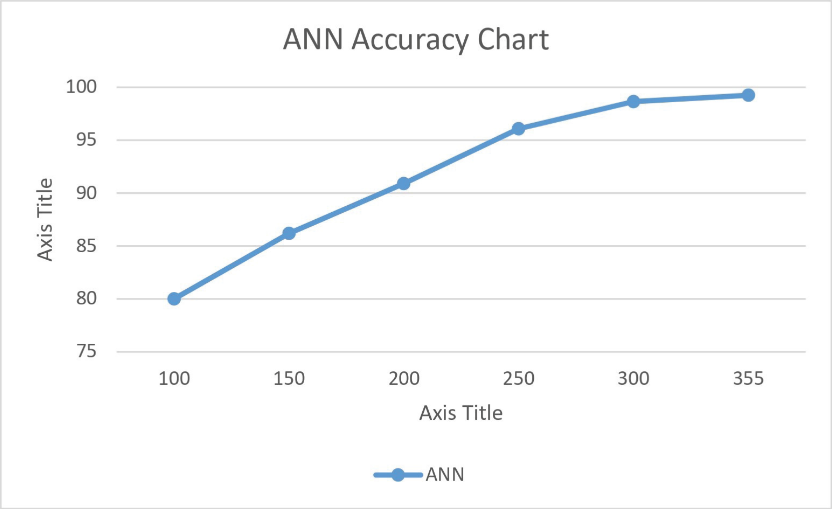


Figure 5.3: Artificial Neural Networks (ANN) Accuracy Chart

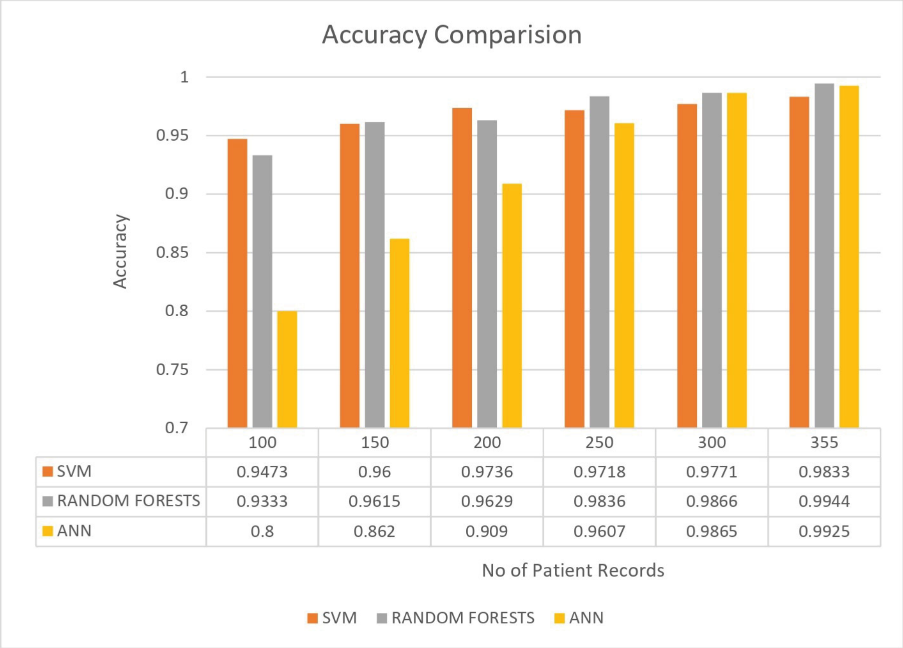
The accuracy of Artificial Neural Networks (ANN) with each record set is represented in Figure 5.3.

5.2.4 Results Comparison

Based on the experiments conducted, the overall accuracy results are tabulated for comparison in Table 5.5. A pictorial representation of performance of each algorithm at different record sets is presented in Figure 5.4.

Number of Patient Records	**Support Vector Machine (SVM) Accuracy**	**Random Forest (RF) Accuracy**	**Artificial Neural Networks (ANN) Accuracy**
100	0.9473	0.9333	0.8%
150	0.96	0.9615	0.862%
200	0.9736	0.9629	0.909%
250	0.9718	0.9836	0.9607%
300	0.9771	0.9866	0.9865%
355	0.9833	0.9944	0.9925%

Table 5.5: Comparison using Performance Metric - accuracy



	100	150	200	250	300	355
■ SVM	0.9473	0.96	0.9736	0.9718	0.9771	0.9833
■ RANDOM FORESTS	0.9333	0.9615	0.9629	0.9836	0.9866	0.9944
■ ANN	0.8	0.862	0.909	0.9607	0.9865	0.9925

Figure 5.4: Performance Comparison Chart

5.2.5 Feature Importance Results

The importance of all the features in the data set are calculated using feature importance experiment conducted through feature_importance package from sklearn python. The calculated values have been represented in Table 5.6. Features in the table are arranged as per the feature values calculated.

The values calculated for feature importance are represented in a pictorial representation in Figure 5.5.

It was identified that the accuracy of the selected machine learning algorithms was not changed while eliminating 3 least important features. After each feature elimination the experiment was re-conducted and the same results are identified.

Feature Number	**Feature Name**	**Feature Value**
18	Chest CT findings - Advances, Absorption	0.155567
13	Fever	0.135102
21	Lymphocyte count	0.133192
4	Respiratory system disease	0.122220
14	Cough	0.094820
24	PCT - Procalcitonin	0.073363
22	Monocyte count	0.058456
20	Neutrophil count	0.056516
3	Age	0.043897
23	CRP - C-reactive protein	0.041181
19	White Blood Cell Count	0.033647
5	Comorbidity	0.012231
6	Fatigue	0.008432
12	Chest tightness	0.006842
2	Clinical Classification	0.004787
7	Cardiovascular and cerebrovascular disease	0.003711
17	Diarrhea	0.002848
1	Gender	0.002814
8	Malignant tumor	0.002763
10	Digestive system disease	0.002654
0	Days from onset of symptoms to hospital admission	0.002272
15	Liver disease	0.001315
16	Endocrine system disease	0.001202
9	Patient Condition	0.000153
11	Renal disease	0.000016

Table 5.6: Feature Importance

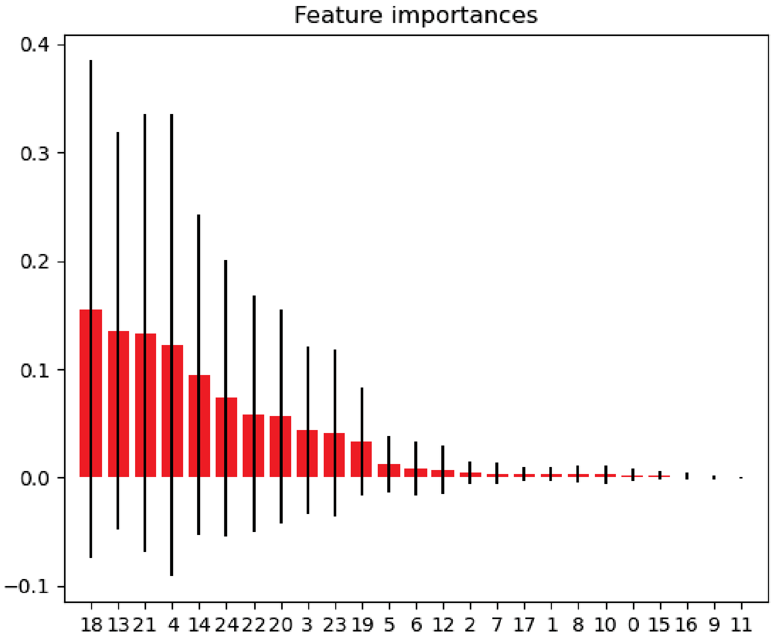


Figure 5.5: Feature Importance Chart

Chapter 6

Analysis and Discussion

6.1 Analysis of Literature Review

According to the results obtained from the Systematic Literature Review (SLR), RQ1 could not be answered thoroughly. In many works, a clear comparison between various machine learning algorithms has been conducted deliberately but the conclusion couldn't be achieved. A comparison model was suggested in [14], [35], [18], [23].

Considering the results from a set of literature, a particular set of algorithms that include: Support Vector Machine (SVM), Artificial Neural Networks (ANNs) and Random Forests (RF) were chosen to perform an experimental evaluation to select the most suitable algorithm to predict COVID-19.

6.2 Analysis of Experiment

The experiment was conducted in 2 phases :

- Evaluation of machine learning algorithms selected from the Literature Review to answer RQ1.
- Feature importance generation for identifying the impact of a particular feature on the prediction of COVID-19 through which RQ2 is answered.

6.2.1 Experiment Phase 1

Quantitative results are analysed with calculated accuracy for each machine learning algorithm to identify the most suitable algorithm for prediction of COVID-19.

- Support Vector Machines (SVMs) showed better results with smaller training data records when compared to other algorithms. There was no much difference observed in the accuracy of prediction when the number of records increased.
- Random Forests (RFs) was found to be the most reliable algorithm among the other algorithms for prediction of COVID-19. Though ruled out by SVMs for smallest number of records, RFs showed consistent growth in accuracy at all stages. RFs has the highest accuracy for classification almost at every record set used.

- Artificial Neural Networks (ANNs) is identified as the most progressive algorithm among the others. In spite of having the lowest accuracy at smaller record sets, ANNs have shown a consistent growth in accuracy levels as the number of records in the dataset increase.

It is observed that Random Forests (RFs) comparatively performs better in terms of accuracy when compared with Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs).

6.2.2 Experiment Phase 2

Experiment Phase 2 is conducted in order to answer RQ2. The aim of this experiment is to identify which features in the dataset influence the predictive result. A descending list of factors that effect the prediction of COVID-19 are tabulated in Table 5.6.

6.3 Discussion

RQ1: Which suitable machine learning technique can be used to predict COVID-19?

By conducting a literature review, several works were considered in connection with the research question and the domain of research. It was concluded that no single algorithm can be marked as the most suitable algorithm. Each technique has its own positives. A set of algorithms were selected which include: Support Vector Machine (SVM), Artificial Neural Networks (ANNs) and Random Forests (RF) were chosen to perform a comparative analysis. For the chosen set of algorithms, accuracy at various stages is analyzed and evaluated. From the results of the experiment, Random Forest (RF) is identified as the suitable machine learning technique that can be used to predict COVID-19.

RQ2: What are the features that will influence the predictive result of COVID-19?

The influence of all the features in the data are calculated by the experiment conducted. The features that show a major change in the prediction are tabulated in Table 6.1. The features that have no affect in the prediction are tabulated in Table 6.2. When features with no affect in the prediction are removed, there was no difference in the accuracy of prediction.

Feature Name	Feature Value
Chest CT findings - Advances, Absorption	0.155567
Fever	0.135102
Lymphocyte count	0.133192
Respiratory system disease	0.122220

Table 6.1: Features that majorly affect the Prediction.

Feature Name	Feature Value
Days from onset of symptoms to hospital admission	0.002272
Liver disease	0.001315
Endocrine system disease	0.001202
Patient Condition	0.000153
Renal disease	0.000016

Table 6.2: Features that have no affect the Prediction.

6.4 Validity Threats

In this section various threats that were identified and mitigated during this research are mentioned.

6.4.1 Internal Validity

One of the internal validity that was identified is the summarising of the literature review. A wrong set of algorithms chosen could change the entire course of the research. To overcome this threat, proper observation was done on the Literature review study in an iterative approach.

6.4.2 External Validity

Improper data pre-processing would affect the results of the experiment, to avoid this the data is checked multiple times after pre-processing. To avoid over fitting, k-fold cross validation has been equipped.

Chapter 7

Conclusions and Future Work

In this research, a systematic literature review has been conducted to identify the suitable algorithm for prediction of COVID-19 in patients. There was no pure evidence found to summarize one algorithm as the suitable technique for prediction. Hence, a set of algorithms which include Support Vector Machine (SVM), Artificial Neural Networks (ANNs) and Random Forests (RF) were chosen. The selected algorithms were trained with the patient clinical information. To evaluate the accuracy of machine learning models, each algorithm is trained with record sets of varying number of patients. Using accuracy performance metric, the trained algorithms were assessed. After result analysis, Random Forest (RF) showed better prediction accuracy in comparison with both Support Vector Machine (SVM) and Artificial Neural Networks (ANNs). The trained algorithms were also assessed to find the features that affect the prediction of COVID-19 in patients.

There is a lot of scope for Machine Learning in Healthcare. For Future work, it is recommended to work on calibrated and ensemble methods that could resolve quirky problems faster with better outcomes than the existing algorithms. Also an AI-based application can be developed using various sensors and features to identify and help diagnose diseases.

As healthcare prediction is an essential field for future, A prediction system that could find the possibility of outbreak of novel diseases that could harm mankind through socio-economic and cultural factor consideration can be developed..

References

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

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

[3] COVID-19 situation reports. Library Catalog: 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.

[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 high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014.

[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.

[26] 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, 2020.

[27] Ilias G Maglogiannis. *Emerging artificial intelligence applications in computer engineering: real word ai systems with applications in ehealth, hci, information retrieval and pervasive technologies*, volume 160. Ios Press, 2007.

[28] Younus Ahmad Malla. A machine learning approach for early prediction of breast cancer. *International Journal of Engineering and Computer Science*, 2017.

[29] Wes McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

[30] Wes McKinney and PD Team. pandas: powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit*, page 1625, 2015.

[31] Sahar A Mokhtar, Alaa Elsayad, et al. Predicting the severity of breast masses with data mining methods. *arXiv preprint arXiv:1305.7057*, 2013.

[32] Ali Narin, Ceren Kaya, and Ziynet Pamuk. Automatic detection of coronavirus disease (covid-19) using x-ray images and deep convolutional neural networks. *arXiv preprint arXiv:2003.10849*, 2020.

[33] Narges Alizadeh Noohi, Marzieh Ahmadzadeh, and M Fardaer. Medical data mining and predictive model for colon cancer survivability. *International Journal of Innovative Research in Engineering & Science*, 2, 2013.

[34] FY Osisanwo, JET Akinsola, O Awodele, JO Hinmikaiye, O Olakanmi, and J Akinjobi. Supervised machine learning algorithms: classification and comparison. *International Journal of Computer Trends and Technology (IJCTT)*, 48(3):128–138, 2017.

[35] Sellappan Palaniappan and Rafiah Awang. Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications*, pages 108–115. IEEE, 2008.

[36] J. Park, K. Kim, and O. Kwon. Comparison of machine learning algorithms to predict psychological wellness indices for ubiquitous healthcare system design. In *Proceedings of the 2014 International Conference on Innovative Design and Manufacturing (ICIDM)*, pages 263–269, 2014.

[37] B. Prabadevi, N. Deepa, K. L. B, and V. Vinod. Analysis of machine learning algorithms on cancer dataset. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pages 1–10, 2020.

[38] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

[39] R Qahwaji and Tufan Colak. Automatic short-term solar flare prediction using machine learning and sunspot associations. *Solar Physics*, 241(1):195–211, 2007.

[40] Yanjun Qi. Random forest for bioinformatics. In *Ensemble machine learning*, pages 307–323. Springer, 2012.

[41] M. Ramaiah, P. Baranwal, S. B. Shastri, M. Vanitha, and C. Vanmathi. Analytical comparison of machine learning techniques for liver dataset. In *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, volume 1, pages 1–5, 2019.

[42] Bendi Venkata Ramana, M Surendra Prasad Babu, NB Venkateswarlu, et al. A critical study of selected classification algorithms for liver disease diagnosis. *International Journal of Database Management Systems*, 3(2):101–114, 2011.

[43] CR Rao and Venkat N Gudivada. *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications.* Elsevier, 2018.

[44] K. Roosa, Y. Lee, R. Luo, A. Kirpich, R. Rothenberg, J. M. Hyman, P. Yan, and G. Chowell. Real-time forecasts of the COVID-19 epidemic in China from February 5th to February 24th, 2020. *Infectious Disease Modelling*, 5:256–263, 2020. Publisher: Elsevier.

[45] Guido Van Rossum et al. Python programming language. In *USENIX annual technical conference*, volume 41, page 36, 2007.

[46] Dawei Wang, Bo Hu, Chang Hu, Fangfang Zhu, Xing Liu, Jing Zhang, Binbin Wang, Hui Xiang, Zhenshun Cheng, Yong Xiong, et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in wuhan, china. *Jama*, 2020.

[47] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, and Bo Xu. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). preprint, Infectious Diseases (except HIV/AIDS), February 2020.

[48] Jenna Wiens and Erica S. Shenoy. Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. *Clinical Infectious Diseases*, 66(1):149–153, 2018. Publisher: Oxford University Press US.

[49] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pages 1–10, 2014.

[50] Zunyou Wu and Jennifer M. McGoogan. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. *Jama*, 323(13):1239–1242, 2020. Publisher: American Medical Association.

[51] Yanyan Xu. Covid19 inpatient cases data.xls url=https://figshare.com/articles/COVID19$_{i}npatient_{c}ases_{d}ata_{x}ls/$12195735/3.

[52] Li Yan, Hai-Tao Zhang, Yang Xiao, Maolin Wang, Chuan Sun, Jing Liang, Shusheng Li, Mingyang Zhang, Yuqi Guo, Ying Xiao, et al. 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*, 2020.

[53] Işık Yilmaz. Comparison of landslide susceptibility mapping methodologies for Koyulhisar, Turkey: conditional probability, logistic regression, artificial neural networks, and support vector machine. *Environmental Earth Sciences*, 61(4):821–836, 2010. Publisher: Springer.

[54] Tony Yiu. Understanding Random Forest, August 2019. Library Catalog: towardsdatascience.com.

[55] Giancarlo Zaccone, Md Rezaul Karim, and Ahmed Menshawy. *Deep Learning with TensorFlow*. Packt Publishing Ltd, 2017.

[56] Victor Zhou. Machine Learning for Beginners: An Introduction to Neural Networks, December 2019. Library Catalog: towardsdatascience.com.

[57] Xiaojin Zhu, John Lafferty, and Ronald Rosenfeld. *Semi-supervised learning with graphs*. PhD Thesis, Carnegie Mellon University, language technologies institute, school of . . . , 2005.

BLEKINGE TEKNISKA HÖGSKOLA · BTH ·

Faculty of Computing, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden