Project Final Report

On

PREDICTION OF COVID-19 INFECTION

Submitted for the requirement of

Project course

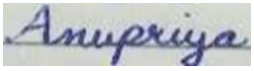
BACHELOR OF ENGINEERING

COMPUTER SCIENCE & ENGINEERING

Submitted to:
SHRUTI BHATLA **(Supervisor)**
E-code: E11963

Submitted By:
Student Group
KAMINI VERMA (20BCS5839)
MUSKAN KUSHWAHA (20BCS5842)
SIDDHARTH PANDEY (20BCS5847)
SOURAV SINGH (20BCS5859)

Co Supervisor Signature
ANUPRIYA
E-code: E10436



DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
CHANDIGARH UNIVERSITY, GHARUAN

June 2022

BONAFIDE CERTIFICATE

Certified that this project report **PREDICTION OF COVID-19 INFECTION** is the bonafide work of Muskan Kushwaha (20BCS5842), Kamini Verma (20BCS5839), Sourav Singh (20BCS5859) and Siddharth Pandey (20BCS5847) who carried out the project work under my/our supervision.

SIGNATURE

Ms. Shruti Bhatla
(SUPERVISOR)
(**CO-SUPERVISOR)**

SIGNATURE

Ms. ANUPRIYA

Submitted for the project viva-voce examination held on 19th May 2022.

INTERNAL EXAMINER
EXAMINER

EXTERNAL

Abstract

Foundation: Over the beyond 4-5 months, the Coronavirus has quickly spread to all regions of the planet. Research is proceeding to track down a remedy for this illness while there is not a great explanation for this episode. As the quantity of cases to test for Coronavirus is expanding quickly step by step, it is difficult to test because of the time and cost factors. Over ongoing years, AI has turned entirely dependable in the clinical field. Utilizing AI to anticipate COVID-19 in patients will lessen the time delay for the aftereffects of the clinical trials and regulate wellbeing laborers to give appropriate clinical treatment to them.

Targets: The fundamental objective of this proposal is to foster an AI model that could foresee whether a patient is experiencing COVID-19. To grow such a model, a writing concentrate close by an analysis is set to recognize an appropriate calculation. To evaluate the elements that influence the expectation model.

Techniques: A Systematic Literature Review is performed to distinguish the most appropriate calculations for the expectation model. Then, at that point, through the discoveries of the writing study, an exploratory model is created for expectation of COVID-19 and to recognize the elements that influence the model.

Results: A bunch of calculations were distinguished from the Literature concentrate on that incorporates SVM (Support Vector Machines), RF (Random Forests), ANN (Artificial Neural Network), which are appropriate for forecast. Execution assessment is directed between the picked calculations to distinguish the method with the most noteworthy precision. Include significance values are produced to recognize their effect on the expectation.

Ends: Prediction of COVID-19 by utilizing Machine Learning could help increment the speed of illness recognizable proof bringing about decreased death rate. Examining the outcomes acquired from tests, Random Forest (RF) was recognized to perform better contrasted with different calculations.

Catchphrases: COVID-19, Machine Learning, Prediction, Supervised Learning, Classification
Strategies

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

Acknowledgement

We have taken endeavors in this venture. Be that as it may, it could never have been conceivable without the caring help and help from numerous people and associations. I might want to stretch out my true because of every one of them.

I might want to communicate our profound and genuine appreciation to our Supervisor and Co-Supervisor Ms. Shruti Bhatla and Ms. Anupriya for permitting us to do the task and giving significant direction all through this examination. Their dynamism, vision and impeccable endeavors have profoundly roused us. They trained us the technique to do the exploration and to introduce the examination function as obviously as could be expected. It was an incredible honor for us to study and work under their direction. I owe the finish of my venture to our undertaking Mentor for her constant help and direction.

My thanks and appreciations likewise go to our educators in directing being developed of the undertaking and individuals who have enthusiastically helped me out with their capacities.

Contents

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

List of Figures

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

List of Tables

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

Chapter 1

INTRODUCTION

Covid are an enormous group of infections that are known to cause disease running from the normal cold to additional serious sicknesses, for example, Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS) [6]. These two illnesses are spread by the Covid 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 most recent infection found in Wuhan, China is called SARS-COV-2 and it causes Covid.

A pneumonia of obscure reason recognized in Wuhan, China was first answered to the World Health Organization (WHO) Country Office in China on 31 December, 2019 [1]. Since, then the quantity of instances of Covid are expanding alongside high loss of life. Covid spread from one city to entire country in only 30 days [50]. On Feb 11, it was named as COVID-19 by World Health Organization (WHO) [5].

As this COVID-19 is spread from one individual to another, Artificial knowledge based electronic gadgets can assume an essential part in forestalling the spread of this infection. As the job of medical services disease transmission specialists has extended, the inescapability of electronic wellbeing information has extended too [13]. The rising accessibility of electronic wellbeing information presents a significant open door in medical care for the two disclosures and useful applications to further develop medical services [48]. This information can be utilized for preparing machine learning calculations to further develop its independent direction as far as foreseeing infections.

As of May 16, 2020, absolutely 44,25,485 instances of COVID-19 have been enlisted and all out number of passings are 3,02,059 [3]. Coronavirus has spread across the globe with around 213 nations and domains impacted [2]. As the ascent in number of instances of tainted Covid immediately dwarfed the accessible clinical assets in medical clinics, come about a significant weight on the medical care frameworks [44]. Because of the restricted accessibility of assets at clinics and the time delay for the aftereffects of the clinical tests, it is

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

what is going on for wellbeing laborers to give legitimate clinical treatment to the patients. As the quantity of cases to test for Covid is expanding quickly step by step, it is absurd to expect to test because of the time and cost factors [25]. In our proposition, we might want to utilize AI strategies to foresee the contamination of Covid in patients.

1.1 Aim

The point of this proposal is to anticipate regardless of whether an individual has COVID 19, utilizing AI procedures. The forecast is performed utilizing the clinical data of the patients. The objective is to distinguish whether a patient might possibly be determined to have COVID-19.

1.2 Objectives

The main objective of our thesis are,

- Distinguishing the most appropriate AI method for forecast, to perform on clinical reports of patients.
- Setting up an AI model that could make exact forecasts of coronavirus in patients.
- Distinguishing the elements that influences the expectation of COVID-19 in patients.

1.3 Research questions

To accomplish the targets of our theory, there are some exploration questions that have been figured out:

1. **Which appropriate AI strategy can be utilized to anticipate coronavirus?**
   Inspiration: The inspiration of the examination question is to direct a conjunctive writing study and examination to see what are the fitting AI calculations that can be best applied to the given information and furthermore to figure out which calculation gives us the best outcomes in foreseeing COVID-19.

2. What are the elements that will impact the prescient consequence of Coronavirus?

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.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

1.4 Defining the extent of the proposition

This exploration centers around advancement of an AI model for anticipating Coronavirus in patients. We additionally work to recognize the elements from the clinical data of patients that would impact the prescient aftereffect of COVID-19. This study doesn't zero in on external factors like climate or any natural variables that could impact results.

1.5 Outline

The postulation structure is separated into various parts which are as per the following:

- Section 1: This part contains the prologue to this theory, point and goals, research questions, and inspiration.
- Section 2: In this part, we examine the foundation of the ideas utilized during the examination.
- Part 3: This section contains the rundown of the works like this theory.
- Section 4: This contains strategies to respond to explore questions. It incorporates exploratory examination like information handling, apparatuses utilized during the analysis, what's more, exploratory arrangement subtleties.
- Section 5: Results got are introduced in this part.
- Part 6: This section comprises of examination and conversations about the outcomes what's more, strategies, the commitment of the proposal to the current exploration, dangers to the legitimacy of the proposal.
- Part 7: In this chapter,we examine the finish of the postulation and conversation on conceivable future work.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

Chapter 3
BACKGROUND

AI is a subset of Artificial Intelligence (AI) and was developed from design acknowledgment where the information can be organized for the comprehension of the clients. As of late, numerous applications have been created utilizing Machine Learning in different fields, for example, medical services, banking, military hardware, space and so on. Presently, AI is a quickly advancing and constantly creating field. It programs PCs utilizing information to advance their exhibition. It learns the boundaries to upgrade the PC programs utilizing the preparation information or its previous encounters.
Utilizing the information, it can likewise anticipate what's to come. AI additionally helps us in building a numerical model utilizing the insights of the information. The primary goal of Machine Learning is that it gains from the feed information with next to no obstruction of people that is, it consequently gains from given data(experience) and gives us the wanted yield where it looks through the patterns/designs in the data [43].
It is extensively arranged into four kinds:
• Supervised Machine Learning.
• Unsupervised Machine Learning.
• Semi-Supervised Machine Learning.
• Reinforcement Machine Learning

Supervised Machine Learning

Managed Learning is a Machine Learning model that is worked to give out forecasts. This calculation is performed by taking a named set of information as info and furthermore referred to reactions as result to gain proficiency with the relapse/characterization model. It creates prescient models from order calculations and relapse procedures.

Classification: predicts discrete reactions. Here, the calculation marks by picking at least two classes for every model. On the off chance that it is done between two classes, it is called two-fold characterization and in the event that it is done between at least two classes, it is called multi-class arrangement. Uses of order incorporates hand composing acknowledgment, clinical imaging and so forth.

Regression: predicts ceaseless reactions. Here, the calculations return a factual esteem. For instance, a bunch of information is gathered with the end goal that individuals are blissful when thought about how much rest. Here, rest and cheerful are the two factors. Presently, the investigation is finished by making predictions [11]. The kinds of famous relapse procedures are:
• Linear regression.
• Logical regression

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

Unsupervised Machine Learning

Dissimilar to the administered realizing, there is no boss here and we just have input information. Here, the essential point is to find specific examples in the information that happen more than others. As indicated by the measurements, it is called thickness assessment. One of the strategies for the thickness assessment is called grouping. Here, the information is shaped into bunches or groupings. Here, the suspicions are made to such an extent that the bunches are found which will coordinate sensibly well with an order. This is an information driven approach that works better when given adequate information. For instance, the films in Netflix.com are proposed in light of the head of grouping of motion pictures where a few comparable motion pictures are gathered in light of client's as of late watched film list.
It for the most part finds the obscure examples in the information however more often than not these approximations are feeble when contrasted and the directed learning [12].

Semi-supervised Machine Learning

The name "semi-administered learning" comes from the way that the information utilized is among managed and unaided learning [57]. Semi-regulated calculation has the propensity to gain both from named and unlabeled information. Semi-administered machine learning gives high precision with a base comment work. Semi-directed AI utilizes for the most part unlabeled information together joined with named information to give better classifiers. As less explanation do whatever is necessary give great exactness, people have less work to do here

Reinforcement Machine Learning

Support gaining gains its way of behaving from an experimentation strategy in a unique climate. Here, the issue is settled by making a fitting move in a specific circumstance to boost the result and to get the obtained results. In Support Learning, there is show of the information or result information. All things considered, whenever the ideal activity is picked, the specialist is quickly informed the award and the next state are not considering the long terms activities. For the specialist to ideally act it ought to have the information about states, rewards, advances and activities effectively.
Formally, the model comprises of [22]:
• a discrete arrangement of climate states, S;
• a discrete arrangement of specialist activities, A;
• a bunch of scalar support signals; commonly {0;1} or the genuine numbers.

2.1 Algorithms

During our examination, we have researched three calculations through which we have performed managed grouping.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

Support Vector Machines (SVM)

Support Vector Machines performs order by developing N-layered hyper plane that isolates the information into two classes [12]. In SVM, the indicator variable is called a characteristic and the changed quality is known as an element. Choosing the most appropriate agent information is called include determination. A bunch of highlights depicting one case is known as a vector.

A definitive objective of SVM displaying is to see as the ideal hyper plane that isolates the groups where on one side of the plane there is target variable and on the opposite side of the plane other classification. The vectors which are close the hyper plane are the help vectors [12]. In Figure 2.1, an average illustration of help vector machine is portrayed.

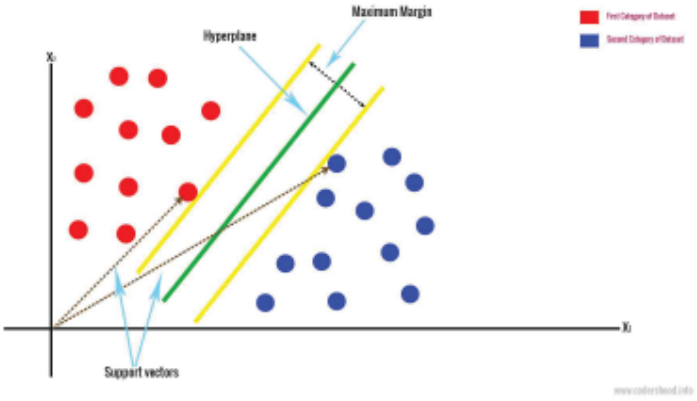


Figure 2.1: Support Vector Machine[7]

Artificial Neural Networks (ANN)

ANNs are an endeavor, in the easiest way, to impersonate the brain arrangement of the human mind [53]. The essential unit of ANN are neurons. A neuron is said to perform capacities on an information and produces a result [56]. Neurons consolidated together are called brain organizations. When the brain networks are framed, preparing of the information is begun to limit the blunder. Eventually, an improving calculation is utilized to further decrease the blunders. The layered design of Artificial Neural Networks (ANNs) is addressed in Figure 2.2.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

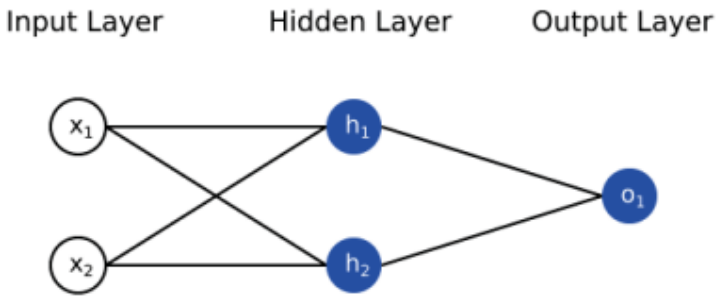


Figure 2.2: Neural Network [56]

Random Forests (RF)

The arbitrary examining and gathering methodologies used in RF empower it to accomplish precise forecasts as well as better speculations [40]. The arbitrary woodlands comprise of huge number of trees. The higher the quantity of uncorrelated trees, the higher the precision [54]. Arbitrary Forest classifiers can help filling a few missing qualities.
Expectation in Random Forests (RFs) is addressed in Figure 2.3

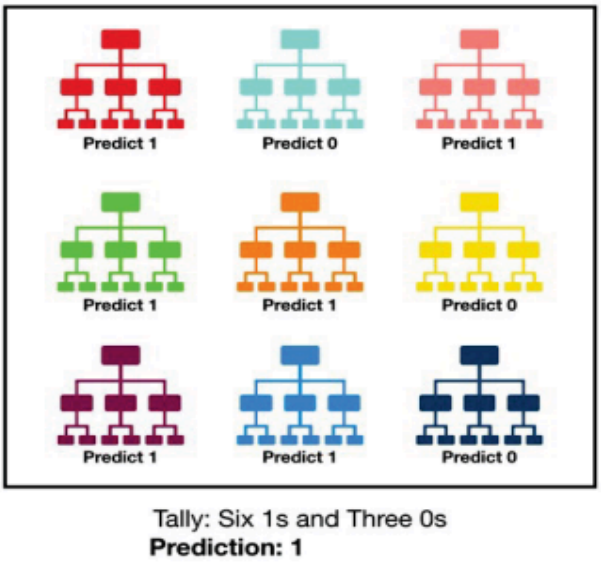


Figure 2.3: Visualization of Random Forest making a prediction. [54]

Chapter 4

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

LITERATURE REVIEW

A methodical writing through the rules of Claes Wohlen [49] and Barbara Kitchenham [24], has been led to examine and answer RQ1. This writing survey centers around the comprehension of a few AI calculations and furthermore recognizing suitable AI calculations that can be utilized for expectation.
There are a few stages that we acted in our exploration, which are:

1. Identifying the key words: We have distinguished the accompanying watchwords which are Supervised Machine Learning calculations, COVID19, characterization, forecast.

2. Formulating the search strings: From the above recognized watchwords, essential catchphrases were chosen to figure out the pursuit string

3. Locating the literature: Utilizing search string, the hunt was performed on different advanced information base stages like Google researcher, IEEE and Science Direct.

4. Following the Inclusion and Exclusion criteria for selection
From the gathered writing, for example, articles and meeting papers, the consideration and avoidance models is carried out to limit our exploration.

Inclusion Criteria

• Papers connected with forecast of COVID-19 utilizing Machine Learning calculations.
• All articles ought to be in English language.

Exclusion Criteria

• Incomplete articles.
• Articles not in English are not considered

5. Evaluating and selecting the literature:
After the execution of the incorporation and rejection rules, further the refining is done through cautious assessment and determination of the assembled writing.

6. Summarizing the literature:
The general discoveries from the assembled writing is summed up and addressed for investigation.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

4.2 Experiment

A trial is directed with the outcomes accomplished from the SLR (Systematic Writing Review) to arrive at the objectives of RQ1 where we recognize the reasonable machine learning method for forecast of COVID-19. The trial is additionally proceeded to construct a model of forecast with the chose calculation to decide RQ2 where the variables that impact the forecast are distinguished

4.2.1 Software Environment

Python

Python is a significant level and compelling general use programming language. It upholds multi-standards. Python has a huge standard library which give devices fit to perform different assignments. Python is a straightforward, less-grouped language with broad elements and libraries. Different programming capacities are used for playing out the analyze in our work. In this proposition, the accompanying python libraries were utilized [45].

• **Pandas** - It is a python bundle that gives expressive information structures intended to work with both social and marked information. It is an open source python library that permits perusing and composing information between information structures [30].
• **Numpy** - It is an open source python bundle for logical figuring. Numpy likewise adds quick exhibit handling abilities to python [29].
• **Tensorflow** - It is a numerical open source python library planned by Google Brain Team for Machine knowledge [55].
• **Sklearn** - It is an open source python AI library intended to work close by Numpy. It highlights different AI calculations for characterization, bunching and relapse.

4.2.2 Dataset

Data Collection

Information assortment was a fundamental and extended process. In any case the field of research, exactness of the information assortment is fundamental to keep up with attachment. As the clinical data of patients was not freely accessible, it was a firm and drawn-out cycle to gather the information. Different Hospitals and Health Institutes in Sweden also, China were drawn closer to get the most dependable information however because of the present circumstance at medical clinics with weighty inflow of patients with COVID-19, we were unable to get admittance to coordinate data. A serious inquiry was directed on different data sets to assemble open source clinical data of patients determined to have COVID-19.

4.2.3 Data Preprocessing

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

Information pre-processing is a significant interaction being developed of AI model. The information gathered is frequently inexactly controlled with out-of-range values, missing values, and so on. Such information can misdirect the consequence of the analysis.

- Ascription of missing qualities - In our information, missing qualities have been dealt with by utilizing straightforward imputer from sklearn python bundle. The missing qualities are supplanted by utilizing mean technique.
- Encoding Categorical Data - We utilized the bundle of OneHotEncoder in python, this bundle handles absolute information by one-hot or sham encoding plan.

4.2.4 Implementation

The examination was directed in the Python IDLE, which is a default incorporated improvement and learning climate for python. The investigation was led in different stages that are referenced underneath:

- After information assortment, the patients information is separated into record sets containing 100 records, 150 records, 200 records, 250 records, 300 records, 355 records separately.
- A 5-overlay cross approval procedure is utilized to randomize the testing informational collection to obtain exact outcomes. Probe each AI calculation is led by 5-overlap cross approval with every one of the record sets.
- The forecast precision of every calculation at each record set is analyzed and assessed for choosing the reasonable calculation for this informational index.
- An element significance try is directed to assess the significance of each quality on the counterfeit characterization task.

4.2.5 Algorithm Configurations

In this part, the setup of the calculations is referenced. Changes made to the design of the calculation can impact the outcomes.

- Support Vector Machines: SVC(kernel = 'direct', random_state = 0)
- Counterfeit 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'))

  Accumulating the ANN:
  ann.compile(optimizer = 'adam', misfortune = 'binary_crossentropy', measurements = ['accuracy'])
- Irregular Forests: RandomForestClassifier(n_estimators = 10, standard = 'entropy', random_state= 0)

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

4.2.6 **Performance Metrics**

It is a fundamental assignment to gauge the presentation of an AI model. As our model requires order, we have involved precision as the presentation metric.

Accuracy

Precision is the measurement utilized in this proposal for assessment of the calculations. It is the most utilized execution metric to assess order procedures. This action permits us to comprehend which model is best at recognizing designs in preparing set to give better expectations in the obscure test informational collection.

$$Accuracy = TP + TN / TP + TN + FP + FN$$

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

Chapter 5

LITERATURE REVIEW RESULT

To answer RQ1, a Systematic Literature Review (SLR) is performed. Which machine learning technique is best for predicting COVID-19? The SLR's goal is to identify the most appropriate algorithms that will aid in the accurate prediction of COVID-19.

Title	Findings

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

Supervised machine learning algorithms: classification and comparison.	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.
Emerging artificial intelligence applications in computer engineering: real word AI systems with applications in E-health.	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.
An empirical comparison of supervised learning algorithms.	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.
Performance evaluation of different machine learning techniques for prediction of heart disease.	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.

Bench marking deep learning models on large healthcare data-sets.	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.
A comparative study of training algorithms for supervised machine learning.	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 .

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

Using machine learning algorithms for breast cancer risk prediction and diagnosis.	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.
Automatic short-term solar flare prediction using machine learning and sunspot associations.	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.
Intelligent heart disease prediction system using data mining techniques.	In this research an 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.

Analysis of cancer data: a data mining approach.	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 9.
Medical data mining and predictive model for colon cancer survivability.	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.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

Analysis of Machine Learning Algorithms on Cancer Dataset.	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.
Analytical Comparison of Machine Learning Techniques for Liver Dataset.	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 other.
Predicting the severity of breast masses with data mining methods.	The article is a comparative study of Decision Tree (DT), Artificial Neural Network (ANN), and Support Vector Machine (SVM) which are analysed on mammographic masses dataset. The results summarize that SVM perform with the highest accuracy followed by ANN.
Data mining applications in healthcare sector: a study.	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.
A Machine Learning Approach for Early Prediction of Breast Cancer.	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.
Comparison of seven algorithms to predict breast Cancer survival contribution to 21 century intelligent technologies and bioinformatics.	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.
A study on classification techniques in data mining.	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.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

A critical study of selected classification algorithms for liver disease diagnosis.	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.
Comparison of machine learning algorithms to predict psychological wellness indices for ubiquitous healthcare system design.	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.
Constructing Inpatient Pressure Injury Prediction Models Using Machine Learning Techniques.	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 .

Table 5.1: Literature Review Results.

The Systematic Literature Review found numerous publications in the healthcare industry that used machine learning techniques (SLR). The majority of the papers compared different machine learning techniques. Healthcare datasets necessitate a comparison of algorithms to determine the best fit. When accuracy is the performance parameter, the most commonly used algorithms are Support Vector Machines (SVMs), Artificial Neural Networks (ANNs), and Random Forests (RFs).

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).

Every one of the above expressed calculations were prepared with the informational index that was gathered and results were deciphered. Execution of every calculation was assessed at

various phases of preparing set. Every calculation was prepared with records establishes containing 100 standards, 150 records ,200 records, 250 records, 300 records, 355 records
separately. This investigation is performed to acquire which calculation would be the
generally reasonable for forecast of COVID-19. Likewise, as the information is parted into more modest sets,
we could likewise asses which calculation would perform better with various datasets
accessible.

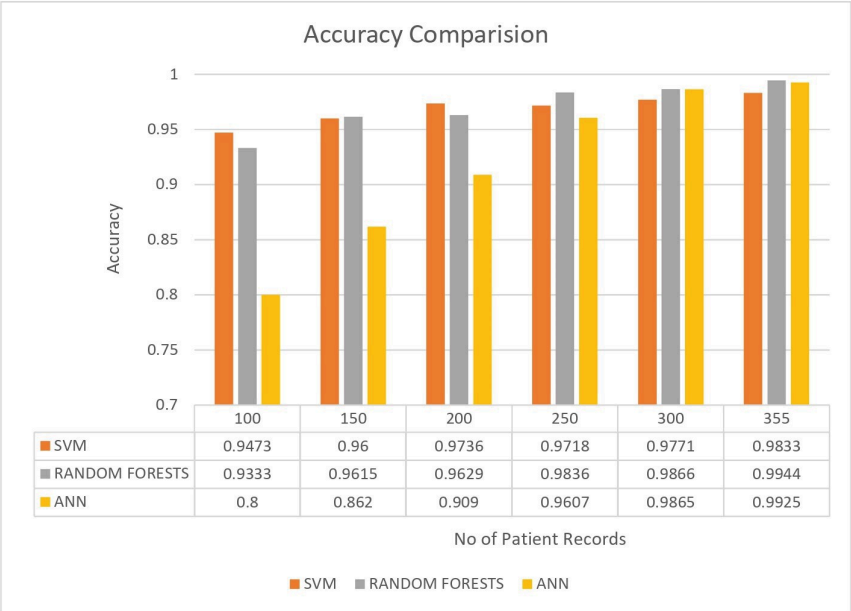
5.2.4 Results Comparison

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

NAME:MUSKAN KUSHWAHA
UID:20BCS5842



	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.

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.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

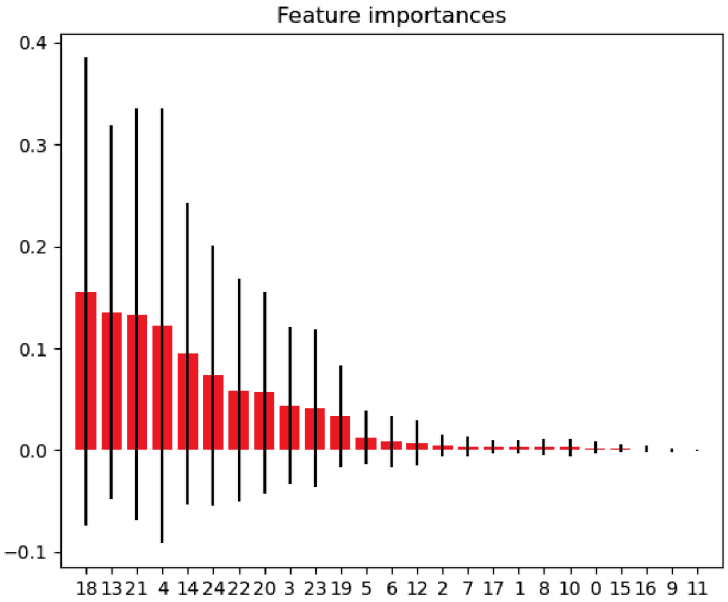


Figure 5.5: Feature Importance Chart

Chapter 6
ANALYSIS AND DISCUSSION

6.1 Analysis of Literature Review

As per the outcomes acquired from the Systematic Literature Review (SLR), RQ1 couldn't be addressed completely. In many works, an unmistakable examination between different AI calculations has been led intentionally however the end couldn't be accomplished. A correlation model was proposed in [14], [35], [18],[23].

Thinking about the outcomes from a bunch of writing, a specific arrangement of calculations that include: Support Vector Machine (SVM), Artificial Neural Networks (ANNs) what's more, Random Forests (RF) were decided to play out a trial assessment to select the most appropriate calculation to anticipate COVID-19.

6.2 Analysis of Experiment

The analysis was directed in 2 stages:

• Assessment of AI calculations chose from the Literature Review to answer RQ1.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

• Include significance age for distinguishing the effect of a specific element on the forecast of COVID-19 through which RQ2 is replied.

6.2.1 Experiment Phase 1

Quantitative outcomes are broken down with determined exactness for each AI calculation to recognize the most reasonable calculation for expectation of COVID-19.

• Support Vector Machines (SVMs) showed improved results with more modest preparation information records when contrasted with different calculations. There was no much contrast seen in the precision of forecast when the quantity of records expanded.

• Arbitrary Forests (RFs) was viewed as the most dependable calculation among the different calculations for expectation of COVID-19. However, precluded by SVMs for most modest number of records, RFs showed steady development in exactness by any means stages. RFs has the most elevated exactness for grouping nearly at each record set utilized.

• Counterfeit Neural Networks (ANNs) is recognized as the most moderate calculation among the others. Regardless of having the most reduced exactness at more modest record sets, ANNs have shown a predictable development in precision levels as the number of records in the dataset increment.

It is seen that Random Forests (RFs) relatively performs better in wording of precision when contrasted and Support Vector Machines (SVMs) and Artificial Brain Networks (ANNs).

6.2.2 Experiment Phase 2

Analyze Phase 2 is directed to answer RQ2. The point of this examination is to distinguish which highlights in the dataset impact the prescient outcome. A slipping rundown of elements that impact the forecast of COVID-19 are classified in Table 5.6.

6.3 Discussion

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

By leading a writing audit, a few works were considered in association with the exploration question and the area of examination. It was reasoned that no single calculation can be set apart as the most appropriate calculation. Every strategy has its own up-sides. A bunch of calculations were chosen which include: Support Vector Machine (SVM), Artificial Neural Networks (ANNs) and Random Forests (RF) were decided to play out a near investigation. For the picked set of calculations, exactness at different stages is examined and assessed. From the consequences of the trial, Irregular Forest (RF) is recognized as the appropriate AI procedure that can be utilized to foresee COVID-19.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

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

The impact of the relative multitude of elements in the information are determined by the examination directed. The highlights that show a significant change in the expectation are organized in Table 6.1. The highlights that have no effect in the expectation are organized in

Table 6.2. Whenever highlights with no effect in the expectation are taken out, there was no distinction in the precision of forecast.

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 part different dangers that were recognized and alleviated during this examination are referenced

6.4.1 Internal Validity

One of the inner legitimacies that was distinguished is the summing up of the writing survey. An off-base arrangement of calculations picked could alter the whole direction of the research. To beat this danger, legitimate perception was done on the Literature survey concentrates in an iterative methodology.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

6.4.2 External Validity

Inappropriate information pre-handling would influence the consequences of the investigation, to stay away from this the information is actually looked at different times after pre-handling. To abstain from over fitting, k-crease cross approval has been prepared.

Chapter 6

CONCLUSIONS AND FUTURE WORK

In this exploration, a precise writing audit has been directed to distinguish the reasonable calculation for forecast of COVID-19 in patients. There was no unadulterated proof found to sum up one calculation as the reasonable strategy for expectation.

Consequently, a bunch of calculations which incorporate Support Vector Machine (SVM), Artificial Brain Networks (ANNs) and Random Forests (RF) were picked. The chose calculations were prepared with the patient clinical data. To assess the precision of AI models, every calculation is prepared with record sets of fluctuating number of patients. Utilizing precision execution metric, the prepared calculations were evaluated.

After outcome investigation, Random Forest (RF) showed better expectation precision in examination with both Support Vector Machine (SVM) and Artificial Neural Networks (ANNs). The prepared calculations were likewise evaluated to track down the elements that influence the expectation of COVID-19 in patients.

There is a ton of degree for Machine Learning in Healthcare. For Future work, it is prescribed to chip away at aligned and troupe techniques that could determine idiosyncratic issues quicker with improved results than the current calculations. Additionally, an Artificial intelligence-based application can be created utilizing different sensors and highlights to distinguish what's more, assist with diagnosing sicknesses.

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

As medical services forecast is a fundamental field for future, An expectation framework that could find the chance of episode of novel infections that could hurt humankind through financial and social component thought can be created.

REFERENCES

1. *WHO EMRO | Questions and answers | COVID-19 | Health topics*

2. *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.*

3. *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, 20.*
4. *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.*

5. *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.*

6. Owusu-Fordjour C., Koomson C., and Hanson D., "The impact of COVID-19 on learning-the perspective of the Ghanaian student," *European Journal of Education Studies*, vol. 7, no. 3, pp. 88–101, 2020.

7. Ting D. S. W., Carin L., Dzau V., and Wong T. Y., "Digital technology and COVID-19," *Nature medicine*, vol. 26, no. 4, pp. 459–461, 2020. pmid:32284618

8. Zambrano-Monserrate M. A., Ruano M. A., and Sanchez-Alcalde L., "Indirect effects of COVID-19 on the environment," *Science of the Total Environment*, vol. 728, p. 138813, 2020. pmid:32334159

9. Zhou C., Su F., Pei T., Zhang A., Du Y., Luo B., et al. "COVID-19: Challenges to GIS with big data," *Geography and Sustainability*, vol. 1, no. 1, pp. 77–87, 2020.

10. Muhammad S., Long X., and Salman M., "COVID-19 pandemic and environmental pollution: a blessing in disguise?" *Science of The Total Environment*, vol. 728, p. 138820, 2020. pmid:32334164

11. Rajkumar R. P., "COVID-19 and mental health: A review of the existing literature," *Asian journal of psychiatry*, vol. 52, p. 102066, 2020. pmid:32302935

12. Millett G. A., Jones A. T., Benkeser D., Baral S., Mercer L., Beyrer C., et al. "Assessing differential impacts of COVID-19 on Black communities," *Annals of Epidemiology*, 2020. pmid:32419766

NAME:MUSKAN KUSHWAHA
UID:20BCS5842

13. Wu JT, Leung K, Leung GM. Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study. Lancet. 2020; 395(10225):689–97.

14. Mohapatra RK, Sarangi AK, Kandi V, Azam M, Tiwari R, Dhama K. Omicron (b. 1.1. 529 variant of sars-cov-2); an emerging threat: current global scenario. J Med Virol. 2021; 2022:1–4.

15. K. R. Bhimala, G. K. PATRA, R. Mopuri, and S. R. Mutheneni, "A deep learning approach for prediction of SARS-CoV-2 cases using the weather factors in India," *Authorea Preprints*, 2020.

16. Arpaci, I., Huang, S., Al-Emran, M. *et al.* Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms. *Multimed Tools Appl* **80,** 11943–11957 (2021)

17. Muhammad, L.J., Algehyne, E.A., Usman, S.S. *et al.* Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset. *SN COMPUT. SCI.* **2,** 11 (2021)

18. Ebubeogu, A.F., Ozigbu, C.E., Maswadi, K. *et al.* Predicting the number of COVID-19 infections and deaths in USA. *Global Health* **18,** 37 (2022)

NAME:MUSKAN KUSHWAHA
UID:20BCS5842