A SMART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS

A Dissertation

Submitted

In Partial Fulfilment of the Requirements for

The Degree of

MASTER OF TECHNOLOGY

In

Computer Science & Engineering

Submitted by:

Md. Ehtisham Farooqui

Under the Supervision of:

Dr. Jameel Ahmad

(Assist. Professor)

Department of Computer Science & Engineering

Faculty of Engineering

INTEGRAL UNIVERSITY, LUCKNOW, INDIA

August, 2020

A SMART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS

A Dissertation

Submitted

In Partial Fulfillment of the Requirements for

The Degree of

MASTER OF TECHNOLOGY

In

Computer Science & Engineering

Submitted by:

Md.Ehtisham Farooqui

Under the Supervision of:

Dr. Jameel Ahmad

(Assist. Professor)

Department of Computer Science & Engineering

Faculty of Engineering

INTEGRAL UNIVERSITY, LUCKNOW, INDIA

Aug, 2020

CERTIFICATE

This is to certify that **Mr. Md. Ehtisham Farooqui** (Enroll. No. 1800102043) has carried out the research work presented in the dissertation titled **"A SMART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS"** submitted for partial fulfillment for the award of the **Master of Technology in Computer Science & Engineering** from **Integral University, Lucknow** under my supervision.

It is also certified that:

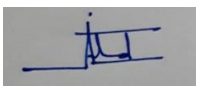
(i) This dissertation embodies the original work of the candidate and has not been earlier submitted elsewhere for the award of any degree/diploma/certificate.

(ii) The candidate has worked under my supervision for the prescribed period.

(iii) The dissertation fulfills the requirements of the norms and standards prescribed by the University Grants Commission and Integral University, Lucknow, India.

(iv) No published work (figure, data, table etc) has been reproduced in the dissertation without express permission of the copyright owner(s).

Therefore, I deem this work fit and recommend for submission for the award of the aforesaid degree.



Dr. Jameel Ahmad
Dissertation Guide
(Assist. Professor)
Department of CSE,
Integral University, Lucknow

Dr. Akheela Khanum
H.O.D.
Department of CSE,
Integral University, Lucknow

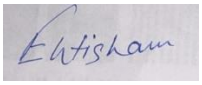
Date: 10-08-2020
Place: Lucknow

DECLARATION

I hereby declare that the dissertation titled **"A SMART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS"** is an authentic record of the research work carried out by me under the supervision of Dr. Jameel Ahmad, Department of Computer Science & Engineering, for the period from August, 2019 to August, 2020 at Integral University, Lucknow. No part of this dissertation has been presented elsewhere for any other degree or diploma earlier.

I declare that I have faithfully acknowledged and referred to the works of other researchers wherever their published works have been cited in the dissertation. I further certify that I have not willfully taken other's work, para, text, data, results, tables, figures etc. reported in the journals, books, magazines, reports, dissertations, theses, etc., or available at web-sites without their permission, and have not included those in this M.Tech dissertation citing as my own work.

Date: 10-08-2020



Signature

Md. Ehtisham Farooqui
Enroll. No. 1800102043

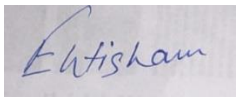
COPYRIGHT TRANSFER CERTIFICATE

Title of the Dissertation: **A SMART DISEASE PREDICTION SYSTEM USING MACHINE LEARNING ALGORITHMS**

Candidate Name: **Md. Ehtisham Farooqui**

The undersigned hereby assigns to Integral University all rights under copyright that may exist in and for the above dissertation, authored by the undersigned and submitted to the University for the Award of the M.Tech degree.

The Candidate may reproduce or authorize others to reproduce material extracted verbatim from the dissertation or derivative of the dissertation for personal and/or publication purpose(s) provided that the source and the University's copyright notices are indicated.



MD. EHTISHAM FAROOQUI

ACKNOWLEDGEMENT

I am highly grateful to the Head of Department of Computer Science and Engineering for giving me proper guidance and advice and facility for the successful completion of my dissertation.

It gives me a great pleasure to express my deep sense of gratitude and indebtedness to my guide **Dr. Jameel Ahmad, Assist. Professor, Department of Computer Science and Engineering,** for his valuable support and encouraging mentality throughout the project. I am highly obliged to him for providing me this opportunity to carry out the ideas and work during my project period and helping me to gain the successful completion of my Project.

I am also highly obliged to **Dr. Akheela Khanum (Associate Professor, Department Of Computer Science and Engineering)** and PG Program Coordinator **Dr. Faiyyaz Ahmad, Assistant Professor, Department of Computer Science and Engineering,** for providing me all the facilities in all activities and for his support and valuable encouragement throughout my project.

My special thanks are going to all of the faculties for encouraging me constantly to work hard in this project. I pay my respect and love to my parents and all other family members and friends for their help and encouragement throughout this course of project work.

Date: 10-08-2020
Place: Lucknow

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

DPS	Disease Prediction System
SVM	Support Vector Machine
MLR	Multilinear Regression
CNN	Convolution Neural Network
ML	Machine Learning

ABSTRACT

Evolution of modern technologies like data science and machine learning has opened the path for healthcare communities and medical institutions, to detect the diseases earliest as possible and it helps to provide better patient care. Accuracy of detecting the possible diseases is reduced when we do not have complete medical data. Furthermore, certain diseases are region-based, which might cause weak disease prediction. Our body shows the symptoms when something wrong is happening within our body, sometime it may be just minor problem but sometimes we can have severe illness and if we do not take care of these symptoms at the early stage then it might be too late to cure the disease. So we are proposing a disease prediction system that can predict the possible diseases based on symptoms so it can be cured at the early stage. It saves time that is required to do the complete diagnosis of the patient and based on the suggestions provided by the system we can only get the patient diagnosed for those diseases that are required. In this paper, we are using machine learning algorithms that try to accurately predict possible diseases. The results generated by the proposed system have an accuracy of up to 87%. The system has incredible potential in anticipating the possible diseases more precisely. The main motive of this study is to help the nontechnical person and freshman doctors to make a correct opinion about the diseases.

CHAPTER 1

INTRODUCTION

1.1 Disease Prediction System:

It is a system that is made by the use of machine learning algorithms for guessing the possible diseases based on the patient's symptoms. The growth of technology has been improving our lives so far. It provides many tools that can save millions of lives, and machine learning is one of them. Machine Learning is used to develop systems that can help us predict so many diseases based on symptoms. It can suggest the doctors, probability of the possible diseases. And diagnosis can be done based on suggestion, thus cost could be reduced.

1.1.1 Some of the existing Disease Prediction System:

- Heart disease prediction system
- Diabetes Disease prediction
- Risk Prediction for Chronic Disease
- Breast cancer prediction

1.1.2 Need for disease prediction system:

Information Technology is being used in each area of human life and improved our lives drastically so far, even Medical Sciences are no exception. A major challenge that every healthcare organization faces is providing medical services at affordable cost, it involves diagnosing patients correctly, and providing effective treatments as a poor diagnosis can lead to fatal consequences that can't be accepted. Hospitals can reduce the costs of diagnosis by using computer-based AI or other decision support systems.

Our technology has evolved so much and can help us to predict many diseases, based on their symptoms, except allergy and respiratory diseases. But to predict the disease we have

to first collect the dataset that can be obtained from hospitals, where they store the data of patients in some sort of patient information system. These systems are used to store huge amounts of patient's data daily that can be represented in the form of numbers, charts, and images. Sadly, searching through such bulk of data is a tedious process thus mostly avoided. The proposed decision-making system will help healthcare organizations to predict the disease based on the current symptoms or previous medical history of a patient. This system will be intelligent enough to guess the possibility of any big disease that can happen to a patient, based on his/her current symptoms.

Some organizations are currently using some sort of decision-making systems but they are not intelligent enough to perform a query to find a disease. They can only perform simple decision-making problems related to the patient's health or disease. Hence this leads to the need for such a decision-making system that can help healthcare organizations to diagnose the exact disease and help the healthcare organizations so that they can provide effective treatment to the patient.

The human body is guarded by the immune system, but sometimes this immune system alone is not capable of preventing our body from diseases. Environmental conditions and living habits of people are the cause of many diseases that are the main reason for a huge number of deaths in the world, and diagnosing these diseases sometimes becomes challenging. There is a need for a precise, consistent, and practicable system to diagnose such diseases in time for proper treatment. With the growth of medical data, many researchers are using these medical data and some machine learning algorithms to help the healthcare communities in the diagnosis of many diseases.

According to McKinsey's [1] report, approx. 50% of Americans have multiple chronic diseases. Because of the living habits of people, the chances of chronic disease is increasing. In India, as the lifestyle of people has improved, the frequency of diseases is also increased. Nearly 62% of death has happened due to non-communicable diseases like heart disorder, cancer, and diabetes. These diseases are often caused by environmental conditions and living habits of people.

Continuous growth in medical data gave us a way to extract the required information to predict the disease. Data mining is applied to detect various types of diseases by using past health data collected from the patient. These disease prediction models are very important to know the presence of disease. For the detection of the disease, we need machine learning techniques (Unsupervised, Semi-Supervised, Reinforcement, Evolutionary, Supervised Learning, and Deep Learning) and raw medical data. This raw dataset can be obtained from famous government hospitals. Machine learning techniques can use the raw data for the learning process and based on that learning later they can predict the disease.

Health is one of the world's challenges for humanity. WHO revealed that for Individual proper health is the essential right. So to keep people fit and healthy proper health care services should be provided. 31 percentage of all deaths worldwide are because of only heart-related problems. Diagnosis and treatment of such disease are very complex, particularly in developing countries, caused by the lack of diagnostic devices and a shortage of doctors and additional resources affecting the correct forecast and treatment of cardiac patients. With this concern in recent times, computer technology and machine learning techniques are used to develop software to assist doctors in deciding heart disease in the

preliminary stage. Early-stage detection of the diseases and predicting the probability of a person to be at risk of any disease can reduce the death rate.

Medical data analysis techniques are used in medical data to extract meaningful patterns and knowledge. Medical information has redundancy, multi-attribution, incompleteness, and a close relationship with time. The problem of using the massive volumes of data effectively becomes a major problem for the health sector. Data analysis provides the methodology and technology to convert these data mounds into useful decision-making information. This prediction system for diseases would facilitate doctors in taking quicker decisions so that more patients can receive treatments within a shorter period, resulting in saving millions of lives.

1.2 The technology used for disease prediction

1.2.1 Machine Learning:

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI [20].

Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming. However, machine learning is not a simple process. As the algorithms

ingest training data, it is then possible to produce more precise models based on that data. A machine-learning model is the output generated when you train your machine-learning algorithm with data. After training, when you provide a model with an input, you will be given an output. For example, a predictive algorithm will create a predictive model [20]. Then, when you provide the predictive model with data, you will receive a prediction based on the data that trained the model.

At a very high level, machine learning is the process of teaching a computer system on how to make accurate predictions when fed data.

Those predictions could be answering whether a piece of fruit in a photo is a banana or an apple, spotting people crossing the road in front of a self-driving car, whether the use of the word book in a sentence relates to a paperback or a hotel reservation, whether an email is a spam, or recognizing speech accurately enough to generate captions for a YouTube video.

The key difference from traditional computer software is that a human developer hasn't written code that instructs the system how to tell the difference between the banana and the apple.

Instead, a machine-learning model has been taught how to reliably discriminate between the fruits by being trained on a large amount of data, in this instance likely a huge number of images labeled as containing banana or an apple.

Machine learning enables models to train on data sets before being deployed. Some machine- learning models are online and continuous. This iterative process of online models leads to an improvement in the types of associations made between data elements. Due to

their complexity and size, these patterns and associations could have easily been overlooked by human observation. After a model has been trained, it can be used in real-time to learn from data. The improvements in accuracy are a result of the training process and automation that are part of machine learning.

Machine learning is enabling computers to tackle tasks that have, until now, only been carried out by people.

From driving cars to translating speech, machine learning is driving an explosion in the capabilities of artificial intelligence - helping software make sense of the messy and unpredictable real world.

Fig. 1.1 Areas in which Machine Learning is used

While machine learning is not a new technique, interest in the field has exploded in recent years.

This resurgence comes on the back of a series of breakthroughs, with deep learning setting new records for accuracy in areas such as speech and language recognition, and computer vision.

What's made these successes possible are primarily two factors, one being the vast quantities of images, speech, video, and text that is accessible to researchers looking to train machine-learning systems.

But even more important is the availability of vast amounts of parallel-processing power, courtesy of modern graphics processing units (GPUs), which can be linked together into clusters to form machine-learning powerhouses.

Today anyone with an internet connection can use these clusters to train machine-learning models, via cloud services provided by firms like Amazon, Google, and Microsoft.

As the use of machine-learning has taken off, so companies are now creating specialized hardware tailored to running and training machine-learning models. An example of one of these custom chips is Google's Tensor Processing Unit (TPU), the latest version of which accelerates the rate at which machine-learning models built using Google's TensorFlow software library can infer information from data, as well as the rate at which they can be trained.

These chips are not just used to train models for Google DeepMind and Google Brain, but also the models that underpin Google Translate and the image recognition in Google Photo,

as well as services that allow the public to build machine learning models using Google's TensorFlow Research Cloud. The second generation of these chips was unveiled at Google's I/O conference in May last year, with an array of these new TPUs able to train a Google machine-learning model used for translation in half the time it would take an array of the top-end GPUs, and the recently announced third-generation TPUs able to accelerate training and inference even further.

As hardware becomes increasingly specialized and machine-learning software frameworks are refined, it's becoming increasingly common for ML tasks to be carried out on consumer-grade phones and computers, rather than in cloud datacenters. In the summer of 2018, Google took a step towards offering the same quality of automated translation on phones that are offline as is available online, by rolling out local neural machine translation for 59 languages to the Google Translate app for iOS and Android.

1.2.2 Approaches to machine learning

Machine learning techniques are required to improve the accuracy of predictive models. Depending on the nature of the problem being addressed, there are different approaches based on the type and volume of the data. Here, we discuss the types of machine learning.

1.2.2.1 Supervised learning

Supervised learning typically begins with an established set of data and a certain understanding of how that data is classified. Supervised learning is intended to find patterns in data that can be applied to an analytics process. This data has labeled features that define

the meaning of data. For example, you can create a machine learning application that distinguishes between millions of animals, based on images and written descriptions.

This approach teaches machines by example.

During training for supervised learning, systems are exposed to large amounts of labeled data, for example, images of handwritten figures annotated to indicate which number they correspond to. Given sufficient examples, a supervised-learning system would learn to recognize the clusters of pixels and shapes associated with each number and eventually be able to recognize handwritten numbers, able to reliably distinguish between the numbers 9 and 4 or 6 and 8.

However, training these systems typically requires huge amounts of labeled data, with some systems needing to be exposed to millions of examples to master a task.

As a result, the datasets used to train these systems can be vast, with Google's Open Images Dataset having about nine million images, its labeled video repository YouTube-8M linking to seven million labeled videos and ImageNet, one of the early databases of this kind, having more than 14 million categorized images. The size of training datasets continues to grow, with Facebook recently announcing it had compiled 3.5 billion images publicly available on Instagram, using hashtags attached to each image as labels. Using one billion of these photos to train an image-recognition system yielded record levels of accuracy -- of 85.4 percent -- on ImageNet's benchmark.

The laborious process of labeling the datasets used in training is often carried out using crowd working services, such as Amazon Mechanical Turk, which provides access to a large

pool of low-cost labor spread across the globe. For instance, ImageNet was put together over two years by nearly 50,000 people, mainly recruited through Amazon Mechanical Turk. However, Facebook's approach of using publicly available data to train systems could provide an alternative way of training systems using billion-strong datasets without the overhead of manual labeling.

1.2.2.2 Unsupervised learning

Unsupervised learning is used when the problem requires a massive amount of unlabeled data. For example, social media applications, such as Twitter, Instagram, and Snapchat, all have large amounts of unlabeled data. Understanding the meaning behind this data requires algorithms that classify the data based on the patterns or clusters it finds. Unsupervised learning conducts an iterative process, analyzing data without human intervention. It is used with email spam detecting technology. There are far too many variables in legitimate and spam emails for an analyst to tag unsolicited bulk email. Instead, machine learning classifiers, based on clustering and association are applied to identify unwanted email.

unsupervised learning tasks algorithms with identifying patterns in data, trying to spot similarities that split that data into categories.

An example might be Airbnb clustering together houses available to rent by neighborhood, or Google News grouping together stories on similar topics each day.

The algorithm isn't designed to single out specific types of data, it simply looks for data that can be grouped by its similarities, or for anomalies that stand out.

1.2.2.3 Reinforcement learning

Reinforcement learning is a behavioral learning model. The algorithm receives feedback from the data analysis, guiding the user to the best outcome. Reinforcement learning differs from other types of supervised learning because the system isn't trained with the sample data set. Rather, the system learns through trial and error. Therefore, a sequence of successful decisions will result in the process being reinforced, because it best solves the problem at hand.

A way to understand reinforcement learning is to think about how someone might learn to play an old school computer game for the first time, when they aren't familiar with the rules or how to control the game. While they may be a complete novice, eventually, by looking at the relationship between the buttons they press, what happens on screen, and their in-game score, their performance will get better over time.

An example of reinforcement learning is Google DeepMind's Deep Q-network, which has beaten humans in a wide range of vintage video games. The system is fed pixels from each game and determines various information about the state of the game, such as the distance between objects on the screen. It then considers how the state of the game and the actions it performs in-game relate to the score it achieves.

Over the process of many cycles of playing the game, eventually, the system builds a model of which actions will maximize the score in which circumstance, for instance, in the case of the video game Breakout, where the paddle should be moved to intercept the ball.

1.2.2.4 Deep learning

Deep learning is a specific method of machine learning that incorporates neural networks in successive layers to learn from data iteratively. Deep learning is especially useful when you're trying to learn patterns from unstructured data. Deep learning complex neural networks are designed to emulate how the human brain works, so computers can be trained to deal with poorly defined abstractions and problems. The average five-year-old child can easily recognize the difference between his teacher's face and the face of the crossing guard. In contrast, the computer must do a lot of work to figure out who is who. Neural networks and deep learning are often used in image recognition, speech, and computer vision applications.

A subset of machine learning is deep learning, where neural networks are expanded into sprawling networks with a huge number of layers that are trained using massive amounts of data. It is these deep neural networks that have fueled the current leap forward in the ability of computers to carry out a task like speech recognition and computer vision.

There are various types of neural networks, with different strengths and weaknesses. Recurrent neural networks are a type of neural net particularly well suited to language processing and speech recognition, while convolutional neural networks are more commonly used in image recognition. The design of neural networks is also evolving, with researchers recently devising a more efficient design for an effective type of deep neural network called long short-term memory or LSTM, allowing it to operate fast enough to be used in on-demand systems like Google Translate.

The AI technique of evolutionary algorithms is even being used to optimize neural networks, thanks to a process called neuro-evolution. The approach was recently showcased by Uber

AI Labs, which released papers on using genetic algorithms to train deep neural networks for reinforcement learning problems.

1.2.3 Evaluation of Machine Learning Models:

Once training of the model is complete, the model is evaluated using the remaining data that wasn't used during training, helping to gauge its real-world performance.

To further improve performance, training parameters can be tuned. An example might be altering the extent to which the "weights" are altered at each step in the training process.

Dataset:

In machine learning, dataset preparation is the process of readying the data for training, testing, and implementation of an algorithm. It’s a multi-step process that involves data collection, cleaning & preprocessing feature engineering, and labeling. These steps play an important role in the overall quality of the machine learning model, as they build on each other to ensure that a model is performing as expected.

Need of Dataset

Dataset is needed to train the machine learning program. Dataset works as input in machine learning programs and then that program learns using that dataset and when learning is done then machine learning programs can be used for classifying, detecting, etc.

Dataset Collection

At the heart of all machine learning projects is data. The nature of this data depends on the project, but it will usually be text, image, video, or audio. Dataset collection is the process

of finding or creating suitable data to use for training a machine learning model. For our model dataset can be collected from different health organizations, hospitals [1].

Dataset Preprocessing

Dataset preprocessing is the act of cleaning and preparing your data for training. This includes organizing and formatting, standardizing, and dealing with missing data. In terms of its importance, many experienced data scientists agree: 80% of their job is data preprocessing.

Data preprocessing is a way to make sure your training data is accurate, complete, and relevant. Sending incomplete or raw data through a model can cause a variety of different errors, which will ultimately result in a much lower overall accuracy.

Feature Engineering

While dataset preprocessing is a way of refining data, feature engineering is the process of creating features to enhance it. Feature engineering allows you to define the most important information in your dataset, and utilize domain expertise to get the most out of it. This might mean breaking data into multiple parts to clarify particular relationships. It might also mean defining features that better represent patterns for your machine learning model [3].

Data Labeling

Data labeling is a key part of data preparation for machine learning because it specifies from which parts of the data, the model will learn from. Though improvements in unsupervised learning have resulted in deep learning projects that do not require labeled data, many machine learning systems still rely on labeled data to learn and perform their given tasks.

Data labeling is often time-consuming and complex. For example, image recognition systems often require bounding boxes drawn around specific objects, while product recommendation and sentiment analysis systems can require complex cultural knowledge for accurate data labeling.

Dataset Quality

In machine learning, the data preparation process leads to the training of your model, so it's important to be thorough [16]. To help put yourself in a strong position for smoother data preparation and model training, make sure you take the time to ensure you have a quality training dataset from the start.

1.2.2 What Is Cloud Computing?

Cloud computing can be described as a virtual pool of shared resources offering compute, storage, database, and network services that can be rapidly deployed at scale.

Fig 1.2 Multiple devices connected to the cloud

Two huge factors have contributed to the success of cloud computing: 1) technological advancements, such as virtualization of compute instances and abundant high-speed internet access, and 2) widespread investment in constantly building and updating infrastructure, which results in economies of scale. Because of these factors, cloud computing can take all the ingredients that make up a traditional data center and makes all these resources available to consumers on an as-needed basis [19].

But what are the types of cloud computing and why is it becoming the new standard? To understand what makes cloud computing successful, you'll need to clearly understand how companies must manage their IT needs and develop products, especially software.

1.2.2.1 The 3 Main Cloud Service Models

Enterprises and consumers can jump into cloud computing in various ways. From the least to the most complicated, the three main cloud service models are as follows:

Software as a Service (SaaS)

Think of something like Gmail, the ubiquitous free webmail service. With a SaaS product, the consumer simply accesses the product through their browser and doesn't have to be concerned with installations or updates. When paid, these services usually are subscription-based.

Platform as a Service (PaaS)

You can think of PaaS offerings as a curated set of services that work together to solve a large business need. For example, a business may want to create a modern microservices-based product, use remote developers, and have the product readily accessible with no delays around the world. A PaaS company will offer a full development environment where the software can be built, tested, and deployed within their predetermined constraints. This frees the customer to focus on the business and creativity of the product, instead of additional concern over infrastructure.

Infrastructure as a Service (IaaS)

IaaS is the public cloud environment at its lowest commoditized levels. The big offerings such as AWS, Azure, and Google Cloud offer their infrastructure's resources, network connectivity, and security compliance as a product that enterprises can use to customize how they see fit to build a cost-optimized software offering.

1.2.2.2 Characteristics of Cloud Technology

Once you are comfortable with the general resources and services that cloud computing offers and how they relate to legacy data centers, you can move on to understanding the three main cloud deployment models.

Public: This the main type of model, with huge offerings such as Amazon Web Service, Microsoft Azure, Google Cloud Platform. These environments are offered to the consumers and are accessible by the public internet. Consumers don't need to be concerned with any infrastructure ownership.

Private: Private cloud is different than a standard on-premise data center. In both private cloud and legacy data centers, the owners need to purchase and manage the resources and employees. However, in a private cloud the environment is designed to have the same resource sharing and scalability as the public cloud, but with improved security, because only the owners can access this particular environment.

Hybrid: Hybrid is a combination of both, with a link over the public internet connecting the private and public clouds. This aids in disaster recovery or situations when the private cloud has reached its limit and needs to leverage the vast resources of a public cloud.

1.2.2.3 Benefits of Cloud Computing

A reduced need for on-site IT staff

No on-site data centers mean no on-site IT staff for the data centers. Cloud service providers have simplified this part of the equation by providing 99.99% uptime Service Level Agreements. However, you will need to have staff that understands how to migrate to the cloud, manage its resources, and contribute to the new DevOps needs in your new deployment.

Cost-effectiveness

It is possible to buy only the cloud services you need and have the option to scale up later when necessary, such as during seasonal peaks. This means you don't have to make huge investments in physical equipment that need maintenance or become obsolete.

Constant Innovation

Cloud offerings are constantly improving to be faster and cheaper. For example, AWS EC2 instances have had many generational changes over the years, and since usage is à la carte (whether fully on-demand or pre-paid for short amounts of time), users can always benefit by upgrading seamlessly to the newest instance types.

Besides, new services are constantly offered such as improved support for Machine Learning or working across different cloud providers. This provides easy access to the latest advancement without large amounts of initial expenditure.

Backup and disaster recovery

Cloud environments offer extensive ways to easily set up backup and disaster recovery which benefit from the user not needing to purchase new infrastructure. You can make your

data redundant across several geographic areas, and you can leverage the different speed and cost options for varying levels of backups to customize your disaster recovery plans.

Shared Responsibility Model

Cloud providers must meet stringent compliance needs to prove they are safe for use by billion-dollar enterprises or governments. Consumers benefit from this emphasis on security, but they must also be aware that they have a part as well. Companies such as AWS will secure the cloud itself, and the customers must secure what is in the cloud — namely their product, using their implementation of the cloud provider's infrastructure.

1.3 Common diseases and their symptoms that can be predicted using DPS

1.3.1. Cholera

Cholera is a serious foodborne illness caused by V. Cholera (classical or El T). It is now more common due to El T or biotype. Most diseases are mild or symptomatic. Epidemics of cholera is a rare occurrence and often creates a major public health problem. They have tremendous potential for rapid spread and death. The epidemic reaches a point and decreases slightly as the viral force decreases. Often, with the introduction of time management measures, the epidemic has already reached its climax and is declining [21].

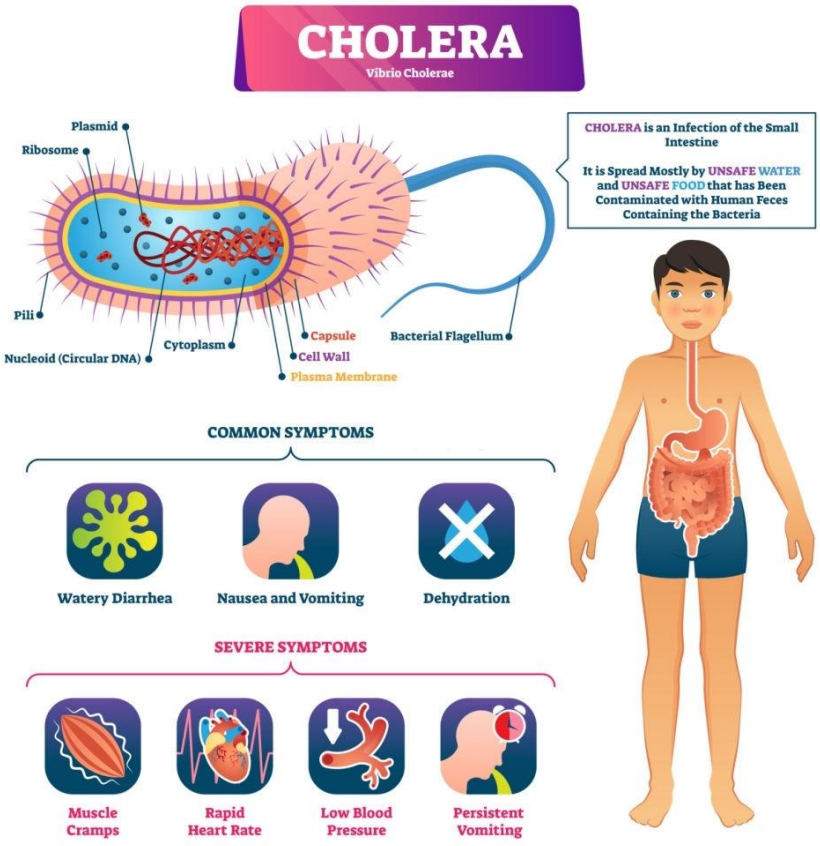


Fig 1.3. symptoms due to Cholera

Symptoms

Most people who get cholera bacterium (Vibrio cholera) are not sick and do not know they have the virus. But because they infect the cholera virus for seven to 14 days, they can still infect others with contaminated water.

Many cholera cases that cause symptoms cause mild or moderate diarrhea which is often difficult to diagnose without diarrhea caused by other complications. Some develop severe signs and symptoms of cholera, usually within a few days of infection.

Symptoms of cholera infection may include:

- Diarrhea. Cholera-related diarrhea suddenly strikes and can quickly cause the loss of a dangerous fluid (about one liter) per hour. Cholera attacks are usually brown, brown-like to the color from which the rice is extracted.
- Nausea and vomiting. Constipation occurs mainly in the early stages of cholera and can last for hours.
- Weight loss. Dryness can develop within a few hours after the onset of symptoms of cholera and can range from mild to severe. A weight loss of 10% or more of body weight indicates body weight.
- Signs and symptoms of cholera include nausea, fatigue, blurred vision, dry mouth, excessive dryness, dry and wrinkled skin on the back when glistening on a stick, with little or no urination, low blood pressure and irregular heartbeat.

Weight loss will be followed by fatigue and constant tiredness. This may happen due to electrolyte imbalance.

Electrolyte imbalance

Electrolyte imbalances can lead to major symptoms and signs such as:

• Muscle cramps: This results in rapid loss of salts such as sodium, chloride, and potassium.

• Repulsion: This is one of the biggest problems with weight loss. It occurs when low blood pressure causes a drop in blood pressure and a decrease in the amount of oxygen in your body. If left untreated, severe hypovolemic shock can cause death within minutes.

1.3.2. Typhoid

Typhoid fever is a systemic illness and severe paralysis. Before the nineteenth century, typhus and typhoid fever were considered such. Enteric fever is a different name for typhoid. Salmonella typhi and paratyphi colonize only humans.

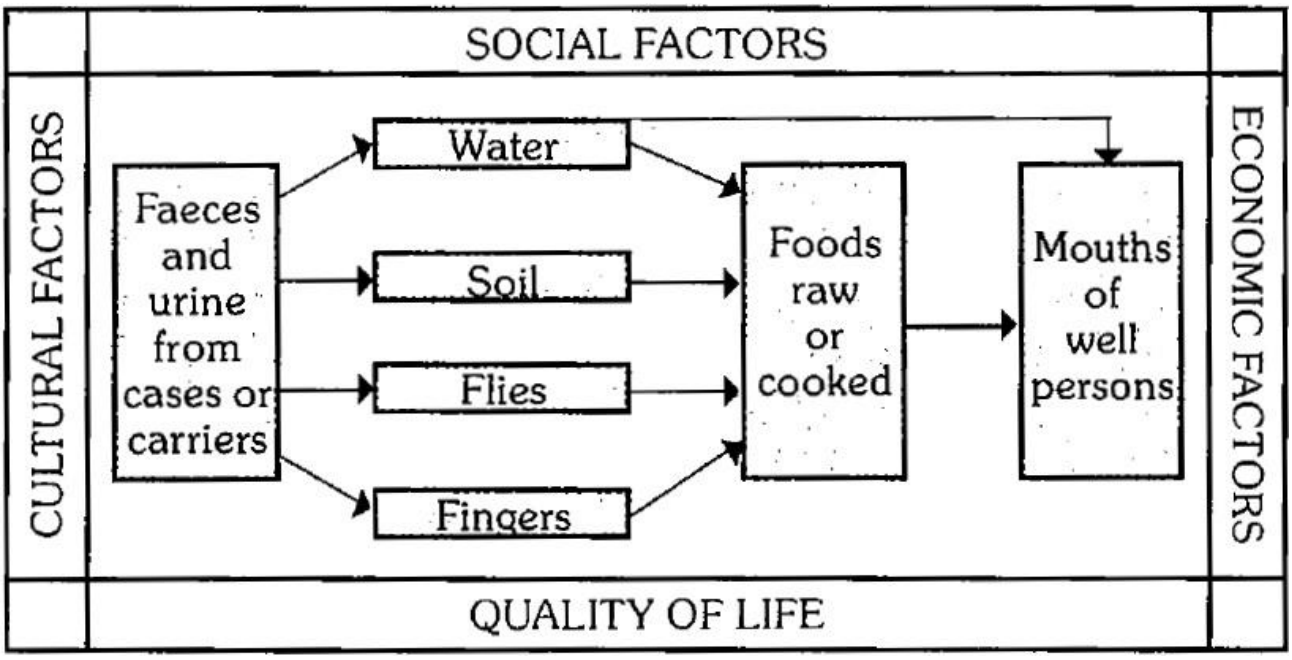


Fig 1.4. Dynamics of typhoid fever transmission

Creatures are acquired through the addition of food or water and are contaminated with human excrement in infected people. Individual deliveries are rare. Typhoid is a global health problem. It is seen in children older than one year.

A typhoid outbreak in developing countries is leading to many deaths. The recent development of antimicrobial drugs raises serious concerns. Typhoid fever is most common in tropical areas. It often occurs in areas, where pollution levels are low. A bacterium called salmonella typhi causes typhoid fever.

Salmonella paratomy can also cause fever and stomach symptoms. This disease caused by both organs is called enteric fever. The disease develops with a common fever, lasting about three to four weeks, severe bradycardia (due to enlargement of the lymph nodes in the abdomen), and constipation.

Worldwide, typhoid fever affects an estimated six million people with over 6,000,000 deaths a year. About 80 percent of all homicides and deaths occur in Asia, with many more in Africa and Latin America. In Asian countries, India probably has a lot of these cases.

Indian Statisticsid typhoid fever is rampant in India. A health study conducted by the Department of Health and Community Development has shown that the rate of decline has dropped from 102 to 2,219 per 1,000,000 people in various parts of the country. A limited study in an urban setting shows 1 percent of children up to the age of 17 suffer from typhoid fever each year.

Carriers of Typhoid Fever typhoid infection are found mainly in people who carry the disease. Carriers are the people who continue to admire salmonella through their urine and

digestion for one year after typhoid fever. The chronic incurable condition increases by about 2 to 5 percent of cases. Organisms in such conditions make the gall bladder their habitat.

Symptoms

Signs and symptoms may develop slowly - they usually appear one or three weeks after exposure to the disease.

Early illness

When signs and symptoms appear, you may experience:

- A fever that starts and goes up every day, maybe up to 104.9 F (40.5 C)
- Headache
- Weakness and fatigue
- Muscle damage
- Sewing
- Dry cough
- Loss of appetite and weight loss
- Abdominal pain
- Diarrhea or constipation
- Running
- Abdominal cramps

Latter illness

If you do not receive treatment, you can become: motionless, dull, and lie with your eyes closed in what is known as Typhoid status.

Besides, life-threatening problems often arise during this time.

For some people, the signs and symptoms may come back for up to two weeks after the fever has lessened.

1.3.3. Jaundice

Jaundice, also known as icterus, is a condition that is characterized by yellowing of the skin and white of the eyes. It is a clinical mark or symptom, not a disease in itself. The yellow color is caused by the amount of bile pigment known as bilirubin in the body. Generally, bilirubin is formed by the breakdown of hemoglobin during the destruction of red blood cells.

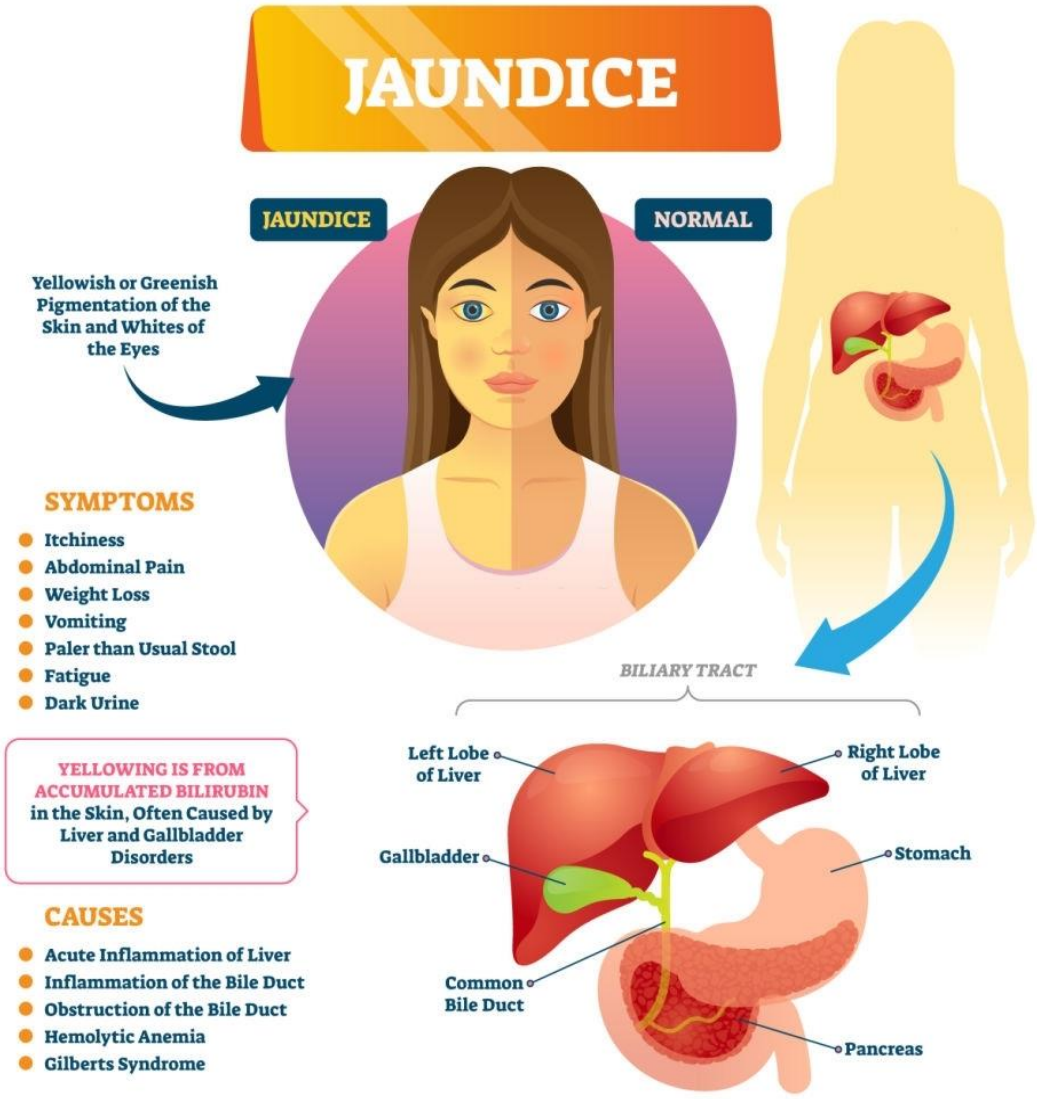


Fig 1.5. Jaundice Symptoms and causes

Symptoms

Common symptoms of jaundice include:

- Yellow patches on the skin and whites of the eyes, usually start at the scalp and spread to the body

- Pale toilet
- Dark urination
- itchiness
- fatigue
- Abdominal pain
- weight loss
- vomiting
- fever

1.3.4. Hepatitis

Hepatitis is an inflammation of the liver. It can be caused by viruses (five different viruses-- called A, B, C, D and E causing Hepatitis Virus), viral infections, or persistent exposure to alcohol, drugs, or toxic chemicals, such as those found in aerosol tests and paint doses, or as a result, of autoimmune disorders.

Hepatitis results in damage to or decreases the ability to perform healthy life-saving functions, including harmful filtration, infectious agents, storing blood sugar and converting it into usable energy forms, and producing much of the protein needed for health.

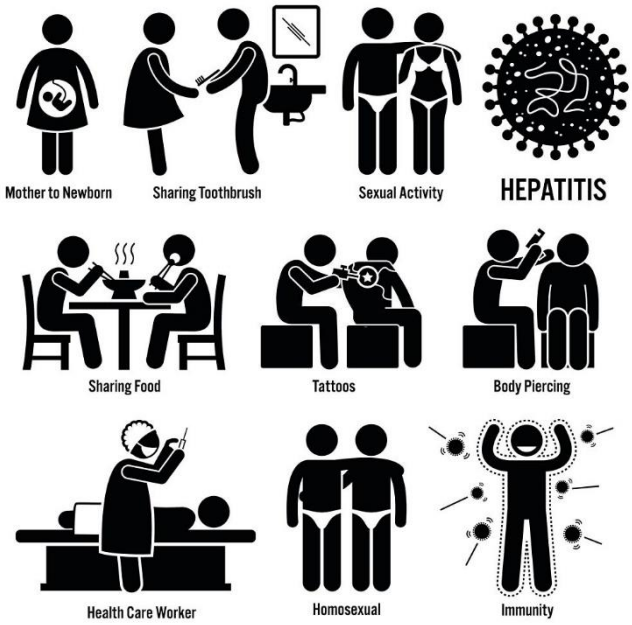


Fig 1.6. Transmission of Hepatitis

The symptoms seen in Hepatitis vary depending on the cause and the overall health of the infected person. However, sometimes, the symptoms can be very mild. The most common manifestations in the clinic are general weakness and fatigue, loss of appetite, nausea, fever, abdominal pain, and tenderness. A key factor is the presence of jaundice (yellowing of the skin and eyes occurs when the liver fails to break down excess bile dye in the blood).

Depending on development and size, Hepatitis can be classified as acute or chronic. In severe Hepatitis, clinical features usually decrease without treatment for a few weeks or months. However, about 5 percent of cases continue to be chronic hepatitis, which can last for years. Chronic hepatitis leads to progressive liver damage and cirrhosis.

Hepatitis A:

Hepatitis A is an autoimmune disease. It is usually transmitted through oral infusions (especially water), but is sometimes transmitted to parents; In many cases, it is similar to the symptoms of mild fever and mild jaundice.

Symptoms

Hepatitis Signs and symptoms usually do not appear until you have been infected for a few weeks. But not all people develop cirrhosis of the liver. If you do, symptoms of hepatitis may include:

- Fatigue
- Nausea and sudden vomiting
- Abdominal pain or discomfort, especially on the right side under your lower ribs (with your liver)
- Brown bowel movements
- Loss of appetite
- Low-grade fever
- Dark urination
- Pain in joint
- Yellow skin and whites of your eyes (jaundice)
- Severe itching

These symptoms may be mild and go away within a few weeks. Sometimes, however, hepatitis A infection leads to a serious illness that lasts for several months.

Hepatitis B:

Hepatitis B is an acute viral disease. It primarily spreads parenterally, but sometimes orally as well. However, the main mode of spread is intimate contact and from mother to newborn. Fever, anorexia, nausea, vomiting are the initial symp-toms, and they soon lead to severe jaundice, urticarial skin lesions, arthritis, etc. Some patients become carriers or even remain chronically ill, even though most patients recover in about three to four months [22].

Symptoms

Signs and symptoms of hepatitis B range from mild to severe. They usually appear about one to four months after you've been infected, although you could see them as early as two weeks post-infection. Some people, usually young children, may not have any symptoms.

Hepatitis B signs and symptoms may include:

- Abdominal pain
- Dark urine
- Fever
- Joint pain
- Loss of appetite
- Nausea and vomiting

- Weakness and fatigue
- Yellowing of your skin and the whites of your eyes (jaundice)

Hepatitis C: Hepatitis C is a viral disease that usually occurs after transfusion or parental drug abuse. It often proceeds in a chronic form that is usually asymptomatic but may include liver cirrhosis.

Symptoms

Long-term infection with the hepatitis C virus is known as chronic hepatitis C. Chronic hepatitis C is usually a "silent" infection for many years, as long as the virus causes considerable damage to the liver to produce signs and symptoms of liver disease.

Signs and symptoms include:

- Easy bleeding
- grow easily
- fatigue
- poor appetite
- Yellow discoloration of skin and eyes (jaundice)
- dark-colored urine
- itchy skin
- The burning buildup in your stomach (ascites)
- Swelling of your feet
- weight loss

- Confusion, drowsiness, and slurred speech (hepatic encephalopathy).
- Spider-like blood vessels (spider angioma) on your skin

Hepatitis D:

Hepatitis D or delta hepatitis is caused by the hepatitis D virus. It usually occurs simultaneously or in the form of superinfection in the case of hepatitis B, thus increasing its severity [22].

Symptoms

Hepatitis D does not always cause symptoms. When symptoms occur, they often include:

- Yellowing of the skin and eyes, known as jaundice.
- joint pain
- stomach ache
- vomiting
- loss of appetite
- dark-colored urine
- fatigue

The symptoms of hepatitis B and hepatitis D are similar, so it can be difficult to determine what disease is causing your symptoms. In some cases, hepatitis D can make hepatitis B

symptoms worse. It can also cause symptoms in people who have hepatitis B but who have never had symptoms.

Hepatitis E:

Hepatitis E is transmitted via oral stool passage; Usually from contaminated water. Chronic infection does not occur but acute infection can be fatal in pregnant women.

Symptoms

Symptoms of hepatitis E can vary. Some people feel so mild without any signs or symptoms that they hardly notice.

Others, however, may experience somewhat different symptoms, usually appearing 15–60 days after exposure to the virus.

Possible symptoms of Hepatitis E include:

- fatigue and general fatigue
- poor appetite
- fever
- Nausea
- vomiting
- Jaundice, or yellowing of the skin and whitening of the eyes
- pain in the upper abdomen, especially on the liver

- light, clay-colored feces
- dark-colored urine

1.3.5. Malaria

Malaria is a deadly disease. It is usually spread by an infected anopheles mosquito bite. Infected mosquitoes carry the Plasmodium parasite. When this mosquito bites you, the parasite escapes into your bloodstream. Once the parasites are inside your body, they travel to the liver, where they mature. After several days, mature parasites enter the bloodstream and begin to infect red blood cells. Within 48 to 72 hours, the parasites inside the red blood cells multiply, causing the infected cells to burst. Parasites continue to infect red blood cells, resulting in symptoms that last two to three days at a time.

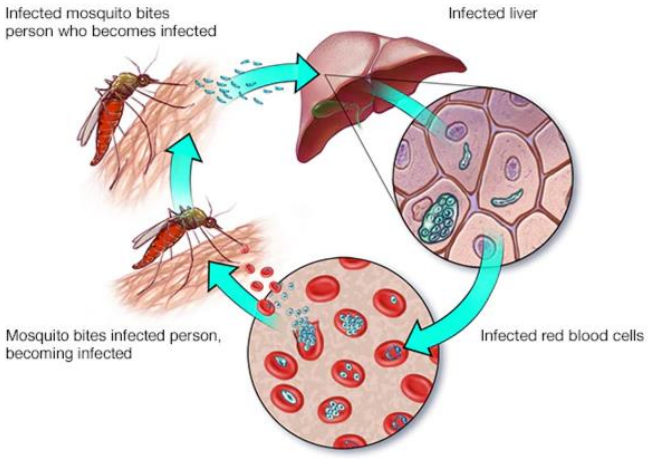


Fig 1.7. How malaria spread

Malaria is a very common disease in developing countries. The word malaria is derived from the word malaria which means bad air. Ronald Ross first discovered the transmission of

malaria by mosquitoes while he was working in India (Secunderabad, AP) in 1897. Malaria is one of the most widespread diseases in the world.

Every year, there are 300 to 500 million clinical cases of malaria in Africa alone. Among all infectious diseases, malaria is the largest contributor to disease burden in terms of deaths and suffering. Malaria kills more than one million children per year in the developing world, accounting for half of the malaria deaths worldwide.

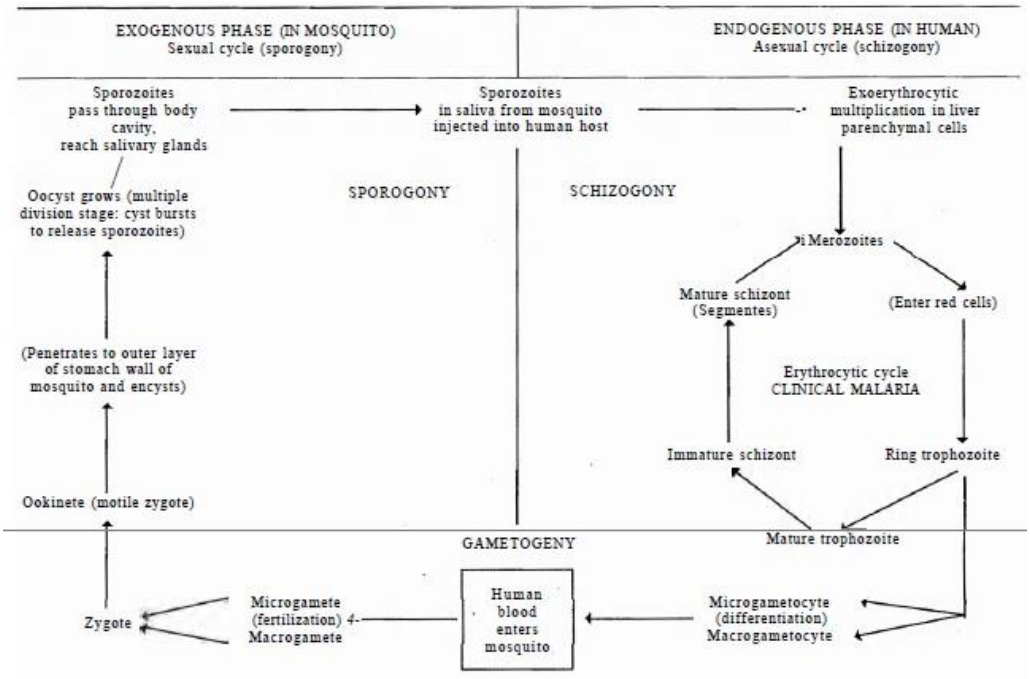


Fig 1.8. The life cycle of malaria

The risk of malaria extends to almost the entire population (about 95 percent) in India. The following states have the highest cases of malaria, including Madhya Pradesh, Maharashtra, Orissa, Karnataka, Rajasthan, Assam, Gujarat, and Andhra Pradesh.

Symptoms:

Symptoms of malaria usually develop within 10 days to 4 weeks after infection. In some cases, symptoms may not develop for several months. Some malarial parasites can enter the body but will remain dormant for a long time.

Common symptoms of malaria include:

- Shaking chills that can range from moderate to severe
- High fever
- profuse sweating
- headache
- Nausea
- vomiting
- stomach ache
- diarrhea
- anemia
- muscle pain
- convulsions
- Coma
- Blood in the stools

1.3.6. Leptospirosis

Leptospirosis is a disease caused by a type of bacteria and is associated with animals. It is more common in tropical countries. The disease is also known as cancer fever; Canicola fever, field-fever, mud fever, seven-day fever, and swineherd disease. Leptospirosis is caused by various strains of bacteria of the genus Leptospira. Of all the varieties that cause disease, Leptospira enterohaemorrhagic is the most severe type.

If not treated properly, it can cause serious complications. Leptospirosis is a disease of animals that can spread to humans. Rats are the most common carrier. Soil contaminated with the urine of infected animals can also cause disease to persons exposed to cattle urine, rat urine, or cattle fluids. Workers exposed to sewage workers, agricultural workers, butchers, meat inspectors, contaminated water, and veterinarians are generally at risk.

Person to person transmission is not possible. Leptospirosis can spread due to contact with urine, blood, or tissues of infected individuals. The organisms enter the body through a pause or mucous membrane in the skin.

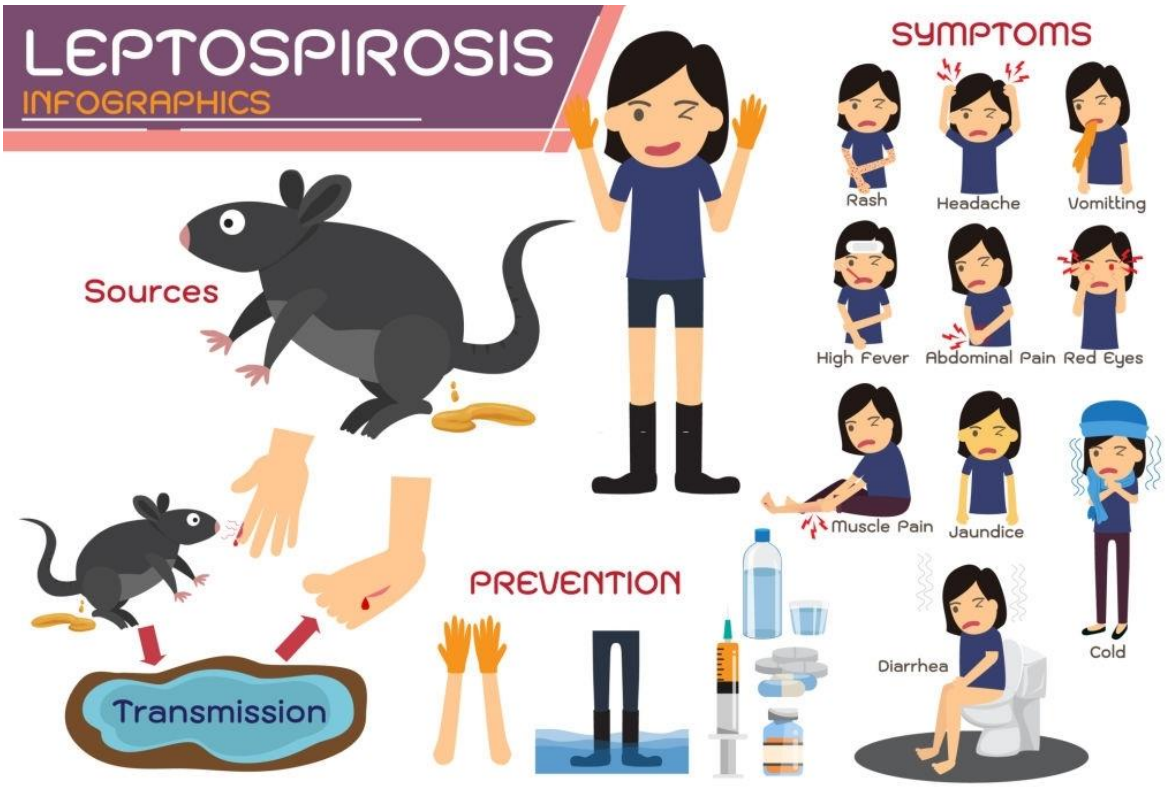


Fig 1.9. Leptospirosis Symptoms and Prevention

Organisms can also be acquired by drinking contaminated water. Infection is usually caused by bathing in contaminated water. Organisms multiply in the blood and tissues of the body. Although the organism can affect any part of the body, the kidneys and liver are usually involved. The incubation period is usually 10 days. This can vary from 2 to 20 days.

Symptoms

In humans, leptospirosis can cause a wide range of symptoms, including:

- High fever
- headache
- feeling cold
- muscle aches
- vomiting
- Jaundice (pale skin and eyes)
- red eyes
- stomach ache
- diarrhea
- Rash

Many of these symptoms can be mistaken for other diseases. Besides, some infected individuals may not have any symptoms.

The time between a person coming into contact with a contaminated source and becoming ill is from 2 days to 4 weeks. The disease usually starts suddenly with fever and other symptoms. Leptospirosis can occur in two stages:

- After the first stage (with fever, chills, headache, muscle aches, vomiting, or diarrhea) the patient may recover for some time but become ill again.
- If the second stage occurs, it is more severe; the person may have kidney or liver failure or meningitis.

The disease lasts from a few days to 3 weeks or longer. Without treatment, recovery may take several months.

1.3.7. Diarrheal Disease

The term 'gastroenteritis' is often used to describe acute diarrhea. Diarrhea is defined as a loose, liquid, or watery stool passage. These liquid feces are usually passed more than three times a day. The attack usually lasts for about 3 to 7 days, but can also last 10 to 14 days.

Diarrhea is a major public health problem in developing countries. Diarrhea causes a huge economic burden on health services. About 15 percent of pediatric beds in India are occupied by admissions due to gastroenteritis. In India, diarrhea disease is a major public health problem in children under 5 years of age. In health institutions, diarrhea accounts for up to one-third of total pediatric admissions.

Diarrhea related disease is an important cause of mortality in children under five years of age. The accident is highest in the age group of 6 to 11 months. The National Diarrheal Disease Control Program has contributed significantly to reduce deaths in children under five years of age.

Symptoms

Signs and symptoms associated with diarrhea may include:

- Loose watery stool
- Cramps in abdominal

- stomach ache
- fever
- Stool bleeding
- mucus in the stool
- swelling
- Nausea
- Urgent need to have a bowel movement

Fig 1.10. Signs and symptoms of diarrhea

1.3.8. Amoebiasis

Amebiasis is an infection caused by a parasite Entamoeba histolytica. The intestinal disease ranges from mild stomach upset and diarrhea to acute abdominal dysentery. Additional intestinal amoebiasis involves the involvement of the liver (liver inflammation), lungs, brain, spleen, skin, etc.

Amebiasis is a common infection of the human gastrointestinal tract. It has a worldwide distribution. It is a major health problem throughout China's southeast and west Asia and Latin America, especially Mexico. It is generally agreed that amoebiasis affects about 15 percent of the Indian population. Amoebiasis has been reported throughout India.

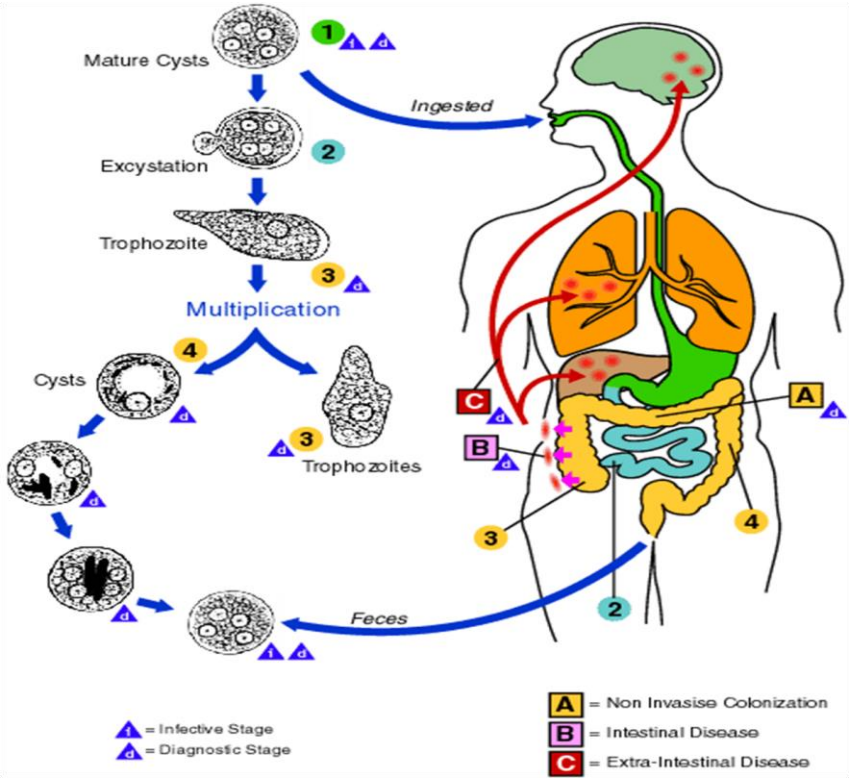


Fig 1.11. The lifecycle of antomeiba histolytica

Symptoms

- Passing the semi-liquid with watery stool about 5-6 times a day
- Presence of blood and excess mucus in the stool
- painful defecation
- Lower abdominal cramps and discomfort, especially during bowel movements
- Passage of sewage and stool
- suppression of urine
- Dehydration (due to loss of body fluids and essential minerals)
- Anemia (due to anemia)
- Cramps in limbs and lower extremities
- Nausea (with or without vomiting)
- Thinning of the anus even after a careful wipe
- Wipe excess water or tissue for general purpose cleaning
- Accidental and involuntary leakage of feces and mucus
- Unknowingly left stains in underwear due to passage of liquid stool
- Stomach upset due to accumulation of gas
- Dry skin (due to loss of water and salt)
- Stool or bowel movements
- sunken eyes and cheeks
- low-grade fever
- Inability to control the passage of feces even by conscious effort
- mood swings and irritability

- dry tongue and mouth
- Weakness and general malaise
- excessive thirst

1.3.9. Brucellosis

Brucellosis is one of the major bacterial zoonoses and is also known in humans as insoluble fever, Malta fever, or Mediterranean fever. It is sometimes spread in humans by direct or indirect contact with infected animals.

The disease can last for several days, months, or sometimes, even years. Brucellosis is both a serious human disease and a disease of animals with serious economic consequences. Brucellosis is a recognized public health hazard found throughout the world.

It is endemic wherever cattle, pigs, goats, and sheep are raised in large numbers. Important endemic areas for brucellosis exist in the Mediterranean, Europe, Central Asia, Mexico, and South America. Animal brucellosis has been reported from practically every state in India. However, there is no statistical information available about the amount of infection in humans in different parts of the country. The prevalence of human brucellosis is difficult to estimate. Many cases inadvertently persist either because they are unclear, or because physicians in many countries are unfamiliar with the disease.

Symptoms

Symptoms of brucellosis can appear anytime from a few days to a few months after you become infected. Signs and symptoms are similar to the flu and include:

- fever
- feeling cold
- loss of appetite
- Sweats
- weakness
- fatigue
- Joints, muscles, and backache
- headache

The symptoms of brucellosis may disappear for weeks or months and then return. Some people have chronic brucellosis and show symptoms even years after treatment. Prolonged signs and symptoms may include fatigue, recurrent fever, arthritis, inflammation of the heart (endocarditis), and spondylitis - inflammatory arthritis that affects the spine and surrounding joints.

1.3.10. Hookworm infection

Hookworm infection is defined as ‘ost ankylostoma’ or any infection caused by the ‘necrosis’. They can occur as single or mixed infections in the same person through various factors, which have to be prevented. Hookworm infection is widely prevalent in India.

Necator americanus is prominent in South India, and Ankylostomes duodenum in North India. Recently, another species from a village near Calcutta, A. Ceylonicum has been reported. The heavily infected areas are found in Assam (tea gardens).

West Bengal, Bihar, Orissa, Andhra Pradesh, Tamil Nadu, Kerala, and Maharashtra. More than 200 million people in India are estimated to be infected. It is believed that 60 to 80 percent of people in West Bengal, Uttar Pradesh, Bihar, Orissa, Punjab, and some areas of Tamil Nadu and Andhra Pradesh are infected with hookworm.

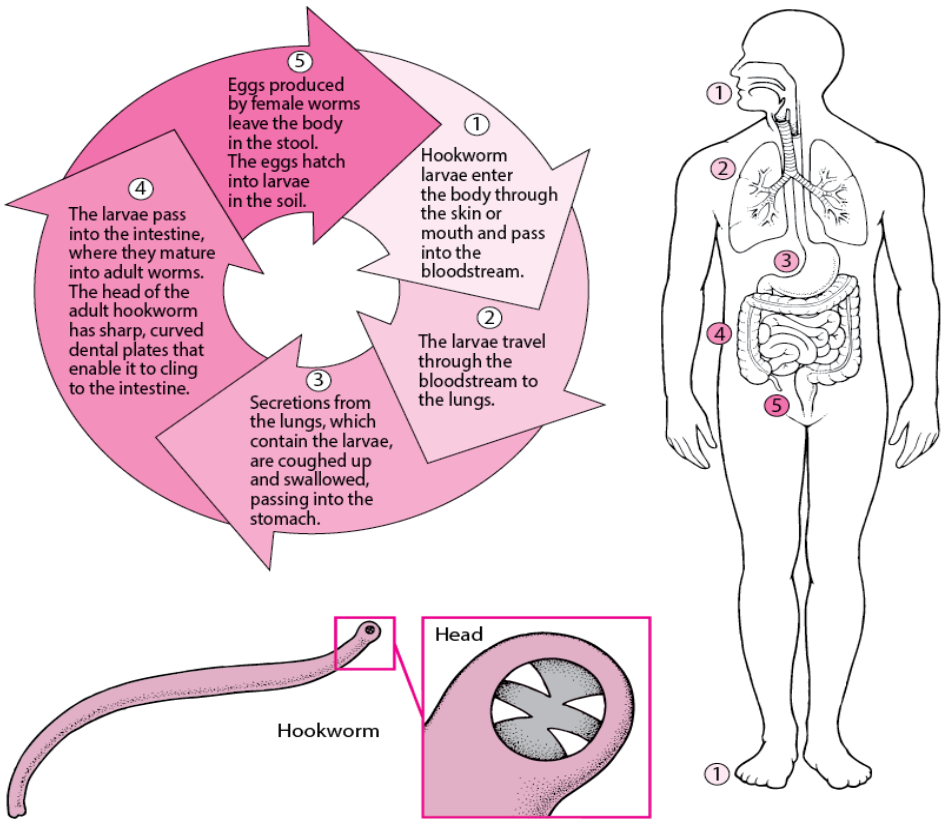


Fig 1.12. Hookworm life cycle

Symptoms

If you are otherwise healthy, you may have no symptoms of infection, the parasite burden is low, and eat lots of iron-rich foods.

If you experience symptoms, they usually start with itching and there is a small rash caused by an allergic reaction in the area that the larvae enter into your skin. This usually occurs after diarrhea as hookworm grows in your intestine. Other symptoms include:

- Stomach ache
- Colic in babies, or cramps and excessive crying
- Intestinal cramps
- Nausea
- Fever
- Blood in your stool
- Loss of appetite
- Itchy rash

1.3.11. Lymphatic filariasis

The main cause of lymphatic filariasis is a parasitic threadworm called Wuchereria bancrofti. These insects originate from mosquito bites from infected people to uninfected people. Various species of mosquitoes are responsible for transporting insects from an infected person to an uninfected one.

According to the WHO report, an estimated 751 million people are at 'risk' for infection, and 120 million have become infected. The public health problems of lymphatic filariasis are the largest in China, India, and Indonesia. These three countries account for about two-thirds of the estimated world total of infected individuals.

The jovial problem in India:

Filaria is a major public health problem in India. There are an estimated six million attacks of acute filarial disease per year, and at least 45 million individuals currently have one or more chronic filarial lesions. The heavily infected areas are found in Uttar Pradesh, Bihar, Andhra Pradesh, Odisha, Tamil Nadu, Kerala, and Gujarat.

The infection is acquired from a person who has filariasis. Maximum infectivity occurs when organisms are circulating in the blood. The largest number appears in the blood at night and retreats from the bloodstream during the day. Their normal habitat is in the lymph nodes.

The mosquito feeds such a person and acquires filaria parasites. Filaria is an infection of an organism when a mosquito bites a person. The parasite is deposited near the site of puncture. It passes through the perforated skin or can penetrate the skin on its own and eventually reach the lymphatic system. Filaria affects all age groups.

Fig 1.13. Filariasis

Causative factors that are compatible with the spread of the disease:

Filaria is mainly seen in developing countries. Lymphatic filariasis is often associated with urbanization, industrialization, illiteracy, poverty, and poor sanitation. The migration of people supported the spread of filariasis. The movement of people from one place to another has expanded into areas of filaria where filaria was not so prevalent. This largely explains the presence of filaria in urban areas of third world countries.

Climate is an important factor in the epidemiology of filariasis. Areas that are moist and moist and the water remains stable throughout the year are a good breeding ground for mosquitoes. It affects the breeding of mosquitoes, their longevity and also determines the development of parasites in the insect vector.

Symptoms

Some of the symptoms of lymphatic filariasis are given below.

• Frequent chills and fever can be a symptom of lymphatic filariasis. Very high fever in the daytime and at night with chills, especially when microfilaria seeps into the bloodstream.

• Inflammation and swelling in the lymph nodes, causing redness, pain, and tenderness in the lymph nodes.

• Symptoms of lymphatic filariasis may include swelling of the penis, arms, legs, and testes in male counterparts, and this is actually when fluid accumulates in the testes that causes a condition called hydrocyl. A female's breasts may be enlarged or swollen, and this occurs when the free flow of lymph fluid through the lymphatic fluid is interrupted by adult worms in their body.

• In some cases, patients with lymphatic filariasis may also have symptoms of inflammation on the genitals, feet, and hands, known as Elephantiasis, which is one of the symptoms of lymphatic filariasis.

• Some people may have a paroxysmal cough at night. This is the type of cough that most asthmatic patients have. It is an allergic reaction that occurs when the microfilaria travels to the lungs through the bloodstream.

• The passing of white urine, known as chyluria, is another characteristic symptom of lymphatic filariasis. This can lead to loss of vital nutrients from the body, which can lead to excretion and weight loss.

1.3.12. Coronavirus disease 2019 (COVID-19)

Coronaviruses are a family of viruses that can cause illnesses such as the common cold, severe acute respiratory syndrome (SARS), and Middle East respiratory syndrome (MERS). In 2019, a new coronavirus was identified as the cause of a disease outbreak that originated in China.

The virus is now known as the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease it causes is called coronavirus disease 2019 (COVID-19). In March 2020, the World Health Organization (WHO) declared the COVID-19 outbreak a pandemic [23].

Public health groups, including the U.S. Centers for Disease Control and Prevention (CDC) and WHO, are monitoring the pandemic and posting updates on their websites. These groups have also issued recommendations for preventing and treating the illness.

Symptoms

Signs and symptoms of coronavirus disease 2019 (COVID-19) may appear in 2 to 14 days after exposure. This time after exposure and before having symptoms is called the incubation period. Common signs and symptoms can include:

- Fever
- Cough
- Tiredness
- Early symptoms of COVID-19 may include a loss of taste or smell.

Other symptoms can include:

- Shortness of breath or difficulty breathing
- Muscle aches
- Chills
- Sore throat
- Runny nose
- Headache
- Chest pain

This list is not all-inclusive. Other less common symptoms have been reported, such as rash, nausea, vomiting, and diarrhea. Children have similar symptoms to adults and generally have mild illness.

The severity of COVID-19 symptoms can range from very mild to severe. Some people may have only a few symptoms, and some people may have no symptoms at all. Some people may experience worsened symptoms, such as worsened shortness of breath and pneumonia, about a week after symptoms start.

Older people have a higher risk of serious illness from COVID-19, and the risk increases with age [23]. People who have existing chronic medical conditions also may have a higher risk of serious illness. Serious medical conditions that increase the risk of serious illness from COVID-19 include:

- Serious heart diseases, such as heart failure, coronary artery disease or cardiomyopathy
- Cancer
- Chronic obstructive pulmonary disease (COPD)
- Type 2 diabetes
- Severe obesity
- Chronic kidney disease
- Sickle cell disease
- A weakened immune system from solid organ transplants

Other conditions may increase the risk of serious illness, such as:

- Asthma
- Liver disease
- Chronic lung diseases such as cystic fibrosis
- Brain and nervous system conditions
- A weakened immune system from bone marrow transplant, HIV or some medications
- Type 1 diabetes
- High blood pressure

This list is not all-inclusive. Other underlying medical conditions may increase the risk of serious illness from COVID-19.

Chapter 2

Security Background

2.1 Dataset security in Disease Prediction System (DPS)

Dataset security is very important in DPS because our system is trained using dataset so we have to secure our dataset. Our system needs a huge amount of data to be trained so we need a true dataset, so our system's performance can be improved [24].

Our dataset security can lie within the 5 V's of big data:

2.1.1 Volume

- Our dataset itself is related to an enormous size.
- Volume is a huge amount of data.
- To determine the value of data, size of data plays a very crucial role. If the volume of data is very large then it can be used to train our system. This means whether a particular data can be considered as a dataset or not, is dependent upon the volume of data.
- Hence while dealing with Dataset it is necessary to consider a characteristic 'Volume'.
- Here we saw how the volume of dataset plays an important role in our system so we need to secure our volume of data.

2.1.2 Velocity

- Velocity refers to the high speed of accumulation of data.

- Enormous and nonstop flow of data. This determines the potential of data that how fast the data is generated and processed to meet the demands.
- Specimen data can help in dealing through the issue like ‘velocity’.
- In our system velocity can be utilized to train our system regularly so that our system can get better using the velocity of medical data every day.
- Security to the velocity of data is needed to feed the system's new data daily.

2.1.3 Variety

- It refers to the nature of data that is structured, semi-structured, and unstructured data.
- It also refers to heterogeneous sources.
- Variety is the arrival of data from new sources that are both inside and outside of a medical institution in our case. It can be structured, semi-structured, and unstructured.
- Structured data: This data is organized data. It generally refers to data that has defined the length and format of data.
- Semi-Structured Data: This data is semi-organized. It is generally a form of data that does not conform to the formal structure of data. Log files are examples of this type of data.
- Unstructured data: This data refers to unorganized data. It generally refers to data that doesn’t fit neatly into the traditional row and column structure of the relational database. Texts, pictures, videos, etc. are examples of unstructured data that can’t be stored in the form of rows and columns.

- Our system needs different kinds of data, structured or unstructured to get trained.
- So we have to ensure that our system securely gets a variety of data [24].

2.1.4 Veracity

- It refers to inconsistencies and uncertainty in data. If it is available it can sometimes get messy and quality and accuracy are difficult to control.
- Multitude of data is variable because dimensions resulting from multiple disparate data types and sources.
- Sometimes when we get the medical information from the hospitals, some of the details are missing or poorly written and it can cause the inconsistencies in the system.
- So managing the veracity in our system is important.

2.1.5 Value

- After taking the 4 Vs there is also one more V which stands for Value. Data that has no Value is of no good to the company, except you turn it into something useful.
- Data alone has importance but it needs to be converted into something valuable to extract Information. Hence, you can state that Value is the most important V of all the 5Vs.
- We can create value by providing relevant disease information to the consumers but we also have to secure consumer information.

2.2 Security background of disease prediction system

This system will be deployed on the cloud so most of the threats that affect the applications that are deployed on the cloud can also be possible threats for our system.

2.2.1 Cloud Security

Cloud security is also called cloud computing security and it states to the restraint and practice of protecting cloud computing environments, applications, data, and information. Cloud security involves safeguarding cloud environments against unauthorized use/access, distributed denial of service (DDOS) attacks, hackers, malware, and other risks. Cloud security provides safety for cloud environments, the related term, cloud-based security, refers to the software as a service (SaaS) delivery model of security services, and it is hosted on the cloud [24].

2.2.2 Security issues associated with the cloud

Cloud computing and storage provide users with capabilities to store and process their data in third-party data centers. Groups use the cloud in a diversity of different service models (with acronyms such as SaaS, PaaS, and IaaS) and deployment models (private, public, hybrid, and community). Security concerns associated with cloud computing fall into two broad categories: security issues faced by cloud providers (organizations providing software, platform, or infrastructure via the cloud) and security issues tackled by their clients

(companies or organizations who host applications or store data on the cloud). The responsibility is shared, however. The cloud provider should guarantee that their structure is protected and that their client's data is protected. While the user must make sure to encrypt their application and use strong passwords.

When a group picks to store data or host applications on the public cloud, it loses its ability to have physical access to the servers hosting its information. As a result, possibly delicate data is at risk from insider attacks. According to a current Cloud Security Association report, insider attacks are the sixth major risks in cloud computing. So, cloud facility providers must guarantee that thorough background checks are directed for employees who have physical access to the servers in the data center. Additionally, data centers must be often monitored for suspicious activity.

To conserve resources, cut costs, and maintain efficiency, cloud service providers often store more than one customers' data on a single server. Consequently, there is a chance that a user's private data may also be viewed by other users (possibly competitors). To handle such sensitive situations, cloud service providers must ensure proper data isolation and logical storage isolation.

The widespread use of virtualization in implementing cloud infrastructure brings unique security concerns for customers or tenants of the public cloud service. Virtualization changes the relationship between the OS and the underlying hardware such as computing, storage, or networking. It introduces an additional layer of virtualization that must itself be properly configured, managed and secured. Specific concerns include virtualization software, or the ability to compromise "hypervisors". While these worries are mostly theoretical, they do

exist. Like a breach in the workstation with the virtualization software can cause the entire data center to sink or to be reconfigured to an attacker's choice.

2.2.3 Security and privacy

2.2.3.1 Identity management: Every enterprise will have its identity management system to control access to information and computing resources. Cloud service providers can join customer's identity management system in their own systems, using multiple techniques like alliance or SSO technology or a biometric-based identification system or provide an identity management system of their own. CloudID, for instance, provides biometric identification. It links the confidential information of the users to their biometrics and stores them in an encrypted fashion. A biometric identification, search encryption technique is performed in the encrypted domain to validate that the cloud provider or potential attackers do not gain entree to any sensitive data or even the data of the individual queries.

2.2.3.2 Physical Security: Cloud service provider IT hardware i.e. physically secure (servers, routers, cables, etc.) against unauthorized access, interference, theft, fire, flood, etc. and ensure that the necessary supplies (such as power) are minimized. Be sufficiently strong to reduce. Probability of Interruption. This can be gained by serving cloud applications from 'world-class' (i.e., professionally specified, designed, built, managed, monitored and maintained) data centers.

2.2.3.3 Personnel security: Security of data worries that are connected to IT and other professionals connected to cloud service provider, handled through para-employment, pre-

employment, and post-employment activities such as security showing potential recruits, security consciousness and training programs, proactive is done.

2.2.3.4 Confidentiality: Providers ensure that all important data (for example credit card numbers) are masked or encrypted and that only authorized users have access to the data in its entirety. In addition, digital identity and credentials must be protected because any data the provider must collect or produce about customer activity in the cloud.

2.2.4 The Three Primary sorts of Cloud Environments Include

2.2.4.1 Public Cloud Services

Hosted by third-party cloud service providers (e.g. Amazon Web Services (AWS), Microsoft Azure, Google Cloud) and usually accessible through web browsers, so identity management, authentication, and access control are essential.

2.2.4.2 Private Clouds

Usually dedicated and accessible to only one organization. However, they're still susceptible to access breaches, social engineering, and other exploits.

2.2.4.3 Hybrid Clouds

Aspects of public and personal clouds are combined, permitting organizations to use more control over their data and resources than during a public cloud environment, yet still, be ready to tap into the scalability and other benefits of the general public cloud when needed.

2.2.5 The Main Cloud Service Models Generally fall under Three Categories

2.2.5.1 Infrastructure as a Service (IaaS)

It enables pre-configured virtualized data center computing resources (i.e. network, storage, and operating systems) for an on-demand model. This will include automating the construction of virtual machines at scale, so it’s serious to think about how virtual machines are provisioned, managed, and spun down.

2.2.5.2 Platform as a Service (PaaS)

Provides tools and other computing infrastructure, enabling organizations to specialize in building and running web applications and services. It mainly support DevOps, developers, operations and teams. Here, management and configuration of self-service entitlements and privileges are key to controlling risk.

2.2.5.3 Software as a Service (SaaS)

Consists of applications hosted by a 3rd party and typically delivered as software services over an internet browser that's accessed on the client’s side. While it removes the need to

deploy and manage applications on end-user devices, possibly any employee can access web services and download content. Thus, proper visibility and access controls are required to watch sorts of SaaS applications accessed, usage, and cost.

2.2.6 What are the Principal Cloud Computing Security Considerations?

2.2.6.1 Lack of Visibility & Shadow IT

Cloud computing makes it easy for anyone to subscribe a SaaS application or maybe to spin up new instances and environments. Users should follow to robust adequate usage policies for gaining authorization for, and for subscribing to, new cloud services or creating new instances.

2.2.6.2 Lack of Control

Leasing a public cloud service means a corporation doesn't have ownership of the hardware, applications, or software on which the cloud services run. make sure that you understand the cloud vendor's approach to those assets.

2.2.6.3 Transmitting & Receiving Data

Cloud applications often interface and assimilate with other databases, services, and applications. This is often typically achieved through an application programming interface (API). It's vital to know the applications and other people who have access to API data and to encrypt any sensitive information.

2.2.6.4 Embedded/Default Credentials & Secrets : Cloud applications may contain embedded and/or default credentials. Default credentials post an increased risk as they'll be

guessable by attackers. Organizations got to manage these credentials as they might other sorts of privileged credentials.

2.2.6.5 Incompatibilities

IT tools architected for on-premise environments or one sort of cloud are frequently incompatible with other cloud environments. Unsuitability can translate into control gaps and visibility that uncover organizations to risk from vulnerabilities, misconfigurations, data leaks, compliance issues and excessive privileged access.

2.2.6.6 Multitenancy

The cloud benefits of shared resources (e.g., lower cost, flexibility, etc.) are the backbone of multitenancy, but it also introduces concerns about data isolation and data privacy.

2.2.6.7 Scalability

Rapid scalability and automation are main benefits of cloud computing, but the flip side is that misconfigurations, vulnerabilities, and other security issues (such as sharing of secrets–APIs, privileged credentials, SSH keys, etc.) also can proliferate at speed and scale. for instance , cloud administrator consoles enable users to swiftly provision, configure, manage, and delete servers on a huge scale. However, each of those virtual machines is born with their own set of privileges and privileged accounts, which require to be properly onboarded and managed. All of this will be further compounded in DevOps environments, which naturally are fast-charging, highly-automated, and have a tendency to treat security as an afterthought.

2.2.6.8 Malware & External Attackers : Attackers can make a living by exploiting cloud vulnerabilities. Security approach (firewalls, vulnerability management, threat analytics, encoding , identity management, etc.) and rapid detection will assist you to scale back risk while leaving you better poised to reply to face up to an attack.

2.2.6.9 Insider Threats – Privileges

Insider related threats (either through negligence or malevolence), generally take the longest to detect and resolve, with the potential to be the foremost harmful. a robust identity and access management framework alongside effective privilege management tools are essential to eliminating these threats and reducing the damage (such as by preventing lateral movement and privilege escalation) once they do occur.

2.2.7 Cybersecurity Threats to Cloud Computing

2.2.7.1 Cryptojacking

Crytojacking may be a fairly new sort of cyberattack, and it's also one which will very easily go under the radar. It centers on the favored practice of mining for cryptocurrencies like Bitcoin. To try to this, you would like computing power, and cybercriminals have found methods of accessing cloud computing systems then using their computing power to mine for cryptocurrency.

Cryptojacking are often very tricky to identify and affect the main issue here is that the incontrovertible fact that when hackers use computing resources from your cloud system means your operation are going to be bogged down , but (crucially) it'll still work. this

suggests that it can seem as if nothing malicious is occurring which perhaps the computers are just battling their processing power.

Many IT teams mistake the symptoms of crypto-jacking as a flaw with an update or a slower internet connection, meaning it takes them for much longer to determine the important problem.

2.2.7.2 Data breaches

Perhaps the foremost common threat to cloud computing is that the issue of leaks or loss of knowledge through data breaches. a knowledge breach typically occurs when a business is attacked by cybercriminals who can gain unauthorized access to the cloud network or utilize programs to look at , copy, and transmit data.

If you employ cloud computing services, a knowledge breach are often extremely damaging, but it can happen relatively easily. Losing data can violate the overall Data Protection Regulation (GDPR), which could cause your business to face heavy fines.

Remember that a knowledge breach can cause many various issues for your business. apart from the fines and loss of knowledge , you'll also lose the trust of your customers, or maybe have your property stolen.

2.2.7.3 Denial of service: One of the foremost damaging threats to cloud computing may be a denial of service (DoS) attacks. These can pack up your cloud services and make them unavailable both to your users and customers, but also to your staff and business as an entire.

Cybercriminals can flood your system with a really great deal of web traffic that your servers aren't ready to deal with . this suggests that the servers won't buffer, and zip are often

accessed. If the entire of your system runs on the cloud, this will then make it impossible for you to manage your business.

2.2.7.4 Insider threats

When we consider cybersecurity challenges, we frequently consider the concept of malicious criminals hacking into our systems and stealing data – however, sometimes the matter originates from the within of the corporate . Recent statistics suggest that insider attacks could account for quite 43 percent of all data breaches.

Insider threats are often malicious – like members of staff going rogue – but they will even be thanks to negligence or just human error. it's important, then, to supply your staff with training, and also make sure that you're tracking the behavior of employees to make sure that they can't commit crimes against the business.

You should also make sure that you've got a correct off-boarding process in situ . This refers to the purpose at which someone leaves the corporate – you would like to make sure that their access to any crucial data is removed which their credentials not add the system. Many businesses get hacked thanks to malicious former employees looking to urge revenge.

2.2.7.5 Hijacking accounts

Perhaps the best threat to a business that uses cloud computing technologies is that the challenge of hijacked accounts. If a criminal can gain access to your system through a staff account, they might potentially have full access to all or any of the knowledge on your servers without you even realizing any crime has taken place. Cybercriminals use techniques like password cracking and phishing emails to realize access to accounts – so once more ,

the key here is to supply your team with the training to know the way to minimize the danger of their account being hijacked.

One of the ways in which your business can minimize the risks involved hijacked accounts, is thru proper permissions management. this suggests that each account across the business should only tend access to the knowledge that they have to try to to their job. this suggests that if an account is hijacked, there's but the criminal can steal.

2.2.7.6 Insecure applications

Sometimes it are often the case that your system is very secure, but you're disappointed by external applications. Third-party services, such applications, can present serious cloud security risks, and you ought to make sure that your team or cyber-security experts take the time to determine whether the appliance is suitable for your network before they need it installed.

Discourage staff from taking matters into their own hands and downloading any application that they think could be useful. Instead, you ought to make it necessary for the IT team to approve any application before it's installed on the system. While this might sound sort of a lengthy step to place in situ, it can effectively deduct the danger of insecure applications.

Of course, it should even be noted here that applications got to be patched whenever possible, so confirm that this is often a neighborhood of the continued role of your IT team.

2.2.8 Cloud Computing Security Best Practices

2.2.8.1 Strategy & Policy: A holistic cloud security program should account for ownership and accountability (internal/external) of cloud security risks, gaps in protection/compliance, and identify controls needed to mature security and reach the specified end state.

2.2.8.2 Network Segmentation

In multi-tenant environments, assess what segmentation is in situ between your resources and people of other customers, also as between your instances. Leverage a zone approach to isolate instances, containers, applications, and full systems from one another when possible.

2.2.8.3 Identity and Access Management and Privileged Access Management

Leverage robust identity management and authentication processes to make sure only authorized users have access to the cloud environment, applications, and data. Least freedom to limit privileged access should be applied and to strengthen cloud resources (for instance, only expose resources to the web as is important, and de-activate unneeded capabilities/features/access). Ensure privileges are role-based, which privileged access is audited and recorded via session monitoring.

2.2.8.4 Discover and Onboard Cloud Instances and Assets

When cloud instances, facilities, and properties are revealed and grouped, then bring them under management (i.e. managing and cycling passwords, etc.). Discovery and onboarding should be automated the maximum amount as possible to eliminate shadow IT.

2.2.8.5 Password Control (Privileged and Non-Privileged Passwords)

Shared passwords utilization should be never allowed. Passwords with other confirmation systems for sensitive areas could be combined. Password management with best practices must be ensured.

2.2.8.6 Vulnerability Management

Regularly perform vulnerability scans and security audits, and therefore the patch is understood vulnerabilities.

2.2.8.7 Encryption

Cloud data must be encrypted everywhere, whether it is at rest or in transit.

2.2.8.8 Disaster Recovery

Be aware of the info backup, retention, and recovery policies and processes for your cloud vendor. Do they meet your internal standards? Does one have break-glass strategies and solutions in place?

2.2.8.9 Monitoring, Alerting, and Reporting

Implement repeated security and user action observing across all environments and instances. Attempt to integrate and centralize data from your cloud provider (if available) with data from in-house and other vendor solutions, so you've got a holistic picture of what's happening in your environment.

CHAPTER 3

LITERATURE REVIEW

3.1 A COMPARATIVE STUDY OF LITERATURE REVIEW ON DISEASE PREDICTION SYSTEM

Table 3.1:

A comparative study of literature review

YEAR	AUTHOR	PURPOSE	TECHNIQUES USED	ACCURACY
2017	MIN CHEN et al, [1]	Disease Prediction by Machine Learning Over Big Data From Healthcare Communities	CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbor, Decision Tree	94.8%
2018	**S**ayali Ambekar et al, [2]	Disease Risk Prediction by Using Convolutional Neural Network	CNN-UDRP algorithm, Naive Bayes and KNN algorithm	The highest accuracy of 82% is achieved by Naïve Bayes.
2015	Naganna Chetty et al, [3]	An Improved Method for Disease Prediction using Fuzzy Approach	KNN classifier, Fuzzy c-means clustering, and Fuzzy KNN classifier	Diabetes: 97.02% Liver disorder: 96.13%

2019	Dhiraj Dahiwade et al, [4]	Designing Disease Prediction Model Using Machine Learning Approach	K-Nearest neighbor (KNN) and Convolutional neural network (CNN)	KNN: 95% CNN: 98%
2017	Lambodar Jena et al, [5]	Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers	Naive Bayes Multilayer Perceptron	95% 99.7%
2016	Dhomse Kanchan B. et al, [6]	Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis	Naive Bayes classification, Decision Tree and Support Vector Machine	Diabetes Disease: 34.89% Heart Disease: 53%
2018	Pahulpreet Singh Kohli et al, [7]	Application of Machine Learning in Disease Prediction	Logistic Regression	Breast Cancer: 95.71% Diabetes: 84.42%

				Heart Disease: 87.12%
			Decision Tree	Breast Cancer: 94.29% Diabetes: 74.03% Heart Disease: 70.97%
			Random Forest	Breast Cancer: 97.14% Diabetes: 81.82% Heart Disease: 77.42%
			Support Vector Machine	Breast Cancer: 97.14% Diabetes: 85.71% Heart Disease: 83,87%
			Adaptive Boosting	Breast Cancer: 98.57%

				Diabetes: 80.52% Heart Disease: 83.87%
2017	Deeraj Shetty et al, [8]	Diabetes Disease Prediction Using Data Mining	Naïve Bayes and KNN	KNN gives better accuracy, compared to Naïve Bayes.
2017	Rashmi G Saboji et al, [9]	A Scalable Solution for Heart Disease Prediction using Classification Mining Technique	Random Forest Algorithm	98%
2019	Rati Shukla et al, [10]	Machine Learning Techniques for Detecting and Predicting Breast Cancer	Naive Bayes Classifier, Logistic Regression, Support Vector Machines (SVM), Artificial Neural Networks and K-Nearest Neighbor	SVM provides a more accurate result compared to others.

2019	Senthilkumar Mohan et al, [11]	Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network and KNN	88.47%
2019	Anjan Nikhil Repaka et al, [12]	Design And Implementing Heart Disease Prediction Using Naive Bayesian	Naïve Bayes	89.77%
2018	Aakash Chauhan et al, [13]	Heart Disease Prediction using Evolutionary Rule Learning	Association Rule	53%
2018	Aditi Gavhane et al, [14]	Prediction of Heart Disease Using Machine Learning	Multi-Layer Perceptron	91%
2015	Ankita Dewan et al, [15]	Prediction of Heart Disease Using a Hybrid Technique	Neural Network, Decision Tree and Naive Bayes	87%

		in Data Mining Classification		

3.2 REVIEW OF TECHNOLOGY USED

3.2.1 Naïve Bayes

Naïve Bayes classifier based on the probabilistic model and depends on the Bayes theorem. In supervised learning, the Naïve Bayes classifier work. The particular features which are described in a class that is not related to other features.

$P(c/y) = P(y/c)* P(C)/ P(y)$

$P(c/y)$ = posterior probability,

$P(c)$ = prior probability of class,

$P(y/c)$ = likelihood probability of the class,

$P(y)$ = prior probability of predictor.

Based on this algorithm, the classification is carried out.

3.2.2 KNN Algorithm

KNN is a classifier that stored all the values of the variable that are recorded and based on those records, the unknown value of the variable is classified. The unknown value is

classified among the similarity of the variable. KNN is a non-parametric classification method. The KNN is divided into two types the first one is structureless NN technique and the second one is structure-based NN technique. The structured based NN in that data is classified into training and testing data.

Common Distance Metrics:

Euclidean distance (continuous distribution):

$$d(a,b) = \sqrt{\Sigma(ai - bi)^2}$$

Hamming distance (overlap metric):

bat (distance = 1)

toned (distance = 3)

Discrete Metric (Boolean metric):

if x = y then d(x,y) = 0. Otherwise, d(x,y) = 1

Determine the class from k nearest neighbor list

bulk vote taken from class labels among the k-nearest neighbors

Weighted factor:

w =1/d (generalized linear interpolation) or 1/d2

3.2.3 CNN Algorithm

For the prediction of risk associated with diseases, we have to perform five steps of algorithm. In the first step, the dataset is converted into the vector form. Then word embedding carried out which adopts zero values to fill the data. The output of word embedding is a convolutional layer. This convolutional layer is taken as input to the pooling layer and then max-pooling operation on the convolutional layer is performed. The pooling layer is connected with the full connected neural network. Lastly, the full connection layer connected to the classifier that is a softmax classifier.

3.2.4 Multilayer Perceptron

It is the most popular network architecture in today's world. Each unit performs a biased weighted sum of their inputs and passes this activation level through a transfer function to produce their output. The units are arranged in a layered feed-forward topology. Simple input-output model that is in network, with the weights and thresholds. Such links can model roles of almost arbitrary complexity, with the number of layers, and the number of units in each layer, determining the function complexity. The important issues in Multilayer Perceptron are the design specification of the number of hidden layers and the number of units in these layers.

Multilayer Perceptron is a nonlinear classifier based on the Perceptron. A Multilayer Perceptron (MLP) is a backpropagation neural network with one or more layers between the input and output layers.

3.2.5 Support Vector Machine

It is one of the supervised learning model that is associated with learning algorithms that analyze data used for classification and regression analysis. It classifies the data points plotted in a multidimensional space into categories by parallel lines called the hyperplane. The classification of data points involves the maximization of a margin between the hyperplane. It can achieve linear classification along with non-linear classification using kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

A SVM is another modest algorithm that each machine learning professional should have in his or her arsenal. In this scenario, the Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. Moreover, support vector machine, abbreviated as SVM can be used for both regression and classification tasks. But it is widely used in classification objectives. Now let's know a bit more about what is support vector machine. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimension space (N – the number of features) that distinctly classify the data points. Now to separate the two classes of data points many possible hyper planes could be chosen. In our objective though we should find a plane that has the maximum margin, for example, the maximum distance between data points of both classes. Maximizing the margin distance providing us with some reinforcement so that future data points can be classified with more confidence. Here hyperplanes are decision boundaries that help classify the data points. Data points falling on either aspect of the hyperplane will be attributed to completely different categories. Here is one more thing we would like to add is that the dimension of the hyperplane depends upon the number of features. If we can find the number of input features is 2 then the hyperplane is just a line. If input features are 3 then the hyperplane

becomes a two-dimensional plane. From here what we can understand is that it becomes very difficult to imagine when the number of a feature exceeds 3.

3.2.6 Adaptive Boosting

Adaptive Boosting (AdaBoost) formulated by Yoav Freund and Robert Schapire is a machine learning algorithm used for classification as well as for regression analysis. It includes the change of a weak classifier into a strong one via the ensemble technique. For this purpose, the prediction of each weak classifier is merged using a weighted average or by taking into account their prediction accuracy as metrics. Initially, all the attributes are given equal weights, then the algorithm assigns a higher weightage to the inaccurate observation. The error is then propagated with every prediction and multiple iterations are done to reduce it until the prediction becomes accurate.

3.2.7 Decision Trees

For training samples of data D, the trees are constructed based on high entropy inputs. These trees are simple and fast constructed in a top-down recursive divide and conquer (DAC) approach. Tree pruning is performed to remove the irrelevant samples on D.

Entropy = $-\sum_{j=1}^{m} p_{ij} log_2 p_{ij}$

3.3 LITERATURE REVIEW

Numerous works have been done related to disease prediction systems using different data mining techniques and machine learning algorithms in medical centers.

MIN CHEN et al, [1] proposed Disease Prediction by Machine Learning Over Big Data From Healthcare Communities and used techniques like CNN-UDRP algorithm, CNN-MDRP algorithm, Naive Bayes, K-Nearest Neighbor, Decision Tree. This proposed system had an accuracy of 94.8%.

Sayali Ambekar et al, [2] recommended Disease Risk Prediction by Using Convolutional Neural Network and used techniques: CNN-UDRP algorithm, Naive Bayes, and KNN algorithm. In this system, structured data is used and it's accuracy reaches 82% and achieved by using Naïve Bayes.

Naganna Chetty et al, [3] developed An Improved Method for Disease Prediction using Fuzzy Approach and used techniques: KNN classifier, Fuzzy c-means clustering, and Fuzzy KNN classifier. In this paper diabetes disease and liver, disorder prediction is done and the accuracy of Diabetes is 97.02% and Liver disorder is 96.13.

Dhiraj Dahiwade et al, [4] suggested Designing Disease Prediction Model Using Machine Learning Approach and used techniques like K-Nearest neighbor (KNN) and Convolutional neural network (CNN). This paper proposed general disease prediction based on symptoms of the patient. The accuracy of KNN is 95% and the accuracy of CNN is 98%.

Lambodar Jena et al, [5] focuses on Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers and used techniques like Naive Bayes and Multilayer

Perceptron. This paper tries to predict Chronic- Kidney-Disease and the accuracy of Naïve Bayes and Multilayer Perceptron is 95% and 99.7% respectively.

Dhomse Kanchan B. et al, [6] proposed the Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis by using techniques like Naive Bayes classification, Decision Tree and Support Vector Machine. The accuracy of this system is 34.89% for Diabetes and 53% for Heart disease.

Pahulpreet Singh Kohli et al, [7] suggested the Application of Machine Learning in Disease Prediction and used techniques like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Adaptive Boosting. This paper focuses on predicting Heart disease, Breast cancer, and Diabetes. The highest accuracies are obtained using Logistic Regression that is 95.71% for Breast cancer, 84.42% for Diabetes, and 87.12% for Heart disease.

Deeraj Shetty et al, [8] developed Diabetes Disease Prediction Using Data Mining by using Naïve Bayes and KNN. This system predicts diabetes and accuracy obtained by KNN is better than Naïve Bayes.

Rashmi G Saboji et al, [9] proposed A Scalable Solution for Heart Disease Prediction using the Classification Mining Technique and used Random Forest Algorithm. This system presents a comparison against Naïve-Bayes classifier but Random Forest gives more accurate results with accuracy 98%.

Rati Shukla et al, [10] recommended Machine Learning Techniques for Detecting and Predicting Breast Cancer and used techniques like Decision Tree, Support Vector Machine,

Random Forest, Naïve Bayes, Neural Network and KNN. In this system, Support Vector Machine give more accurate results than all other algorithms.

Senthilkumar Mohan et al, [11] focuses on Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques and used algorithms like Decision Tree, Support Vector Machine, Random Forest, Naïve Bayes, Neural Network and KNN. The accuracy of this system is 88.47%.

Anjan Nikhil Repaka et al, [12] developed Design and Implementing Heart Disease Prediction Using Naive Bayesian. Any user can use this system using any smartphone device and get the prediction results. The accuracy of this system is 89.77%.

Aakash Chauhan et al, [13] proposed Heart Disease Prediction using Evolutionary Rule Learning. Association Rule is used in this proposed system. This system is not very efficient because it has an accuracy of 53%.

Aditi Gavhane et al, [14] suggested Prediction of Heart Disease Using Machine Learning. Multi-Layer Perceptron model is used in this system. This system predicts heart disease based on basic symptoms like age, sex, pulse rate, etc. The accuracy of this suggested system is 91%.

Ankita Dewan et al, [15] recommended Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification. This system is using techniques like Neural Network, Decision Tree, and Naive Bayes. The accuracy of this system is 87%.

3.4 Proposed Work

3.4.1 Methodology

i. First, a dataset for blood samples and symptoms is collected.

ii. Then a data for diseases related to symptoms are created.

iii. Support Vector Machine is used for classification.

iv. CNN is used for prediction.

v. Decision Tree is used to make successful choices.

3.4.2 Algorithms

In machine learning, we can use different algorithms otherwise known as classifiers to help us predict the disease for our project. Here in our project, we are looking forward to predicting the number of patients that have a disease and the number of patients that do not have the disease running three algorithms to our data set. The reason we are going to use three is that it will allow us to get a better and more reliable predictions. Because if we are using one algorithm or classifier and do not have anything else to compare it with then we cannot say that it is a reliable prediction because it might be giving us a very good accuracy but this algorithm might not be the best or more appropriate one to use for our scenario. Whereas if we use more than one algorithm or classifier in our case two of them, we can compare them with one another and if we find one classifier is giving us an accuracy that is not even in the approximate of the other algorithm provided accuracy we can understand

that something is going wrong. It can be that the algorithm itself is not suitable for the job or we made a mistake in our coding. That is why using one or more algorithm is essential for any prediction based system. Now the algorithms that we have chosen to use in our project are 1. SVM (Support Vector Machine) and 2. MLR (Multilinear Regression). We will be discussing each of those algorithms below.

3.4.2.1 Support Vector Machine (SVM)

Before we start discussing SVM (support vector machine) we need to be accustomed to Linear regression and Logistic regression algorithms. If not it is suggested to look at them before moving on to support vector machine. It is another simple algorithm that every machine learning expert should have in his or her arsenal. In this scenario, the Support vector machine is highly preferred by many as it produces significant accuracy with less computation power. Moreover, support vector machine abbreviated as SVM can be used for both regression and classification tasks. But it is widely used in classification objectives. Now let's know a bit more about what is support vector machine. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimension space (N – the number of features) that distinctly classify the data points. Now to separate the two classes of data points, many possible hyperplanes could be chosen. In our objective though we should find a plane that has the maximum margin, for example, the maximum distance between data points of both classes. Maximizing the margin distance providing us with some reinforcement so that future data points can be classified with more confidence. Here hyperplanes are decision boundaries that help classify the data points. Data points falling on either aspect of the hyperplane will be attributed to completely different categories. Here is

one more thing we would like to add is that the dimension of the hyperplane depends upon the number of features. If we can find the number of input features is 2 then the hyperplane is just a line. If input features are 3 then the hyperplane becomes a two-dimensional plane. From here what we can understand is that it becomes very difficult to imagine when the number of a feature exceeds 3.

We should always remember that support vectors are data points that are close to the hyperplane and influence the position and orientation of the hyperplane. We maximize the margin of the classifier using these support vectors. Now if we delete the support vectors it will change the position of the hyperplane. These are the points that will eventually help us build our SVM. Now we are going to talk a bit about the large margin intuition. To start we can say that in logistic regression we take the output of the linear function and squash the value within the range of [0, 1] using the sigmoid function. If the squashed value is greater than a threshold value (.5) we assign it as label 1, else we assign it a label 0. In SVM, we take the output of the linear function and if that output is greater than 1, we identify it with one class and if the output is -1, we identify it with another class. Since the threshold values are changed to 1 and -1 in SVM, by doing this we can obtain this reinforcement range of values ([-1, 1]) which acts as margin. Next, we are going to talk about cost function and gradient updates.

To start, in the SVM algorithm, we are looking to maximize the margin between the data points and the hyperplane. The loss operates that helps maximize the margin is hinge loss. The 0 is the cost if the predicted value and the actual value are of the same sign. If they're not, we then calculate the loss value. We conjointly add a regularization parameter the value

operates. The objective of the regularization parameter is to balance the margin maximization and loss. Now that we have the loss function, we take partial derivatives concerning the weights to find the gradients. Using the gradients, we can update our weights. Now when there is no misclassification, we can consider our Model correctly predicting the class of our data point, we only have to update the gradient from the regularization parameter. And if the situation does occur when there is indeed a misclassification. For example, our model makes a mistake on the prediction of the class of our data point, we include the loss along with the regularization parameter to perform gradient update. Features are extracted from the test data and forecast the values. We acquire the predictions and compare them with the particular values and print the accuracy of our model. There are another straightforward thanks to implementing the SVM formula. We can use the 'scikit' learn library and simply decided the connected functions to implement the SVM model. The number of lines of code reduces significantly too few lines. At the end of the day, we can conclude by saying that the Support vector machine is an elegant and powerful algorithm. We should use it wisely.

3.4.2.2 MLR (Multilinear Regression)

Simple Linear Regression, where a single Independent/Predictor(X) variable is used to model the response variable (Y). But there may be many situations in which the reply variable is affected by multiple forecaster variables; for such cases, we use the MLR algorithm.

Multilinear Regression is an expansion of Simple Linear regression because it takes multiple forecaster variables to predict the reply variable [25].

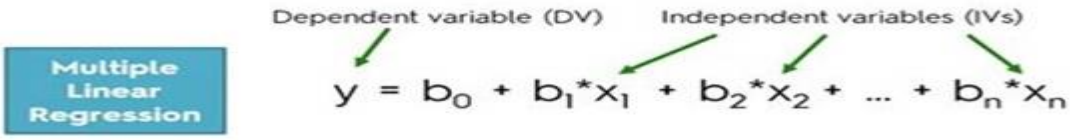
“Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and multiple independent variables.”

Key facts about Multilinear Regression (MLR):

- MLR, the reliant variable(Y) must be continuous or real, but the forecaster or independent variable may be in the categorical or continuous form.
- Each feature variable must model the linear affiliation with the dependent variable.
- MLR tries to fit a regression line through a multidimensional space of data-points.

MLR equation:

In Multilinear Regression, the target variable(Y) is a linear combination of multiple predictor variables x1, x2, x3, ...,xn. Since it is an enhancement of Simple Linear Regression, so the same is applied for the multiple linear regression equation, the equation becomes [25]:



Where,

Y= Output/Response variable

b0, b1, b2, b3 , bn....= Coefficients of the model.

x1, x2, x3, x4,...= Various Independent/feature variable

Assumptions for Multilinear Regression:

- A linear relationship should exist between the Target and predictor variables.
- The regression residuals must be normally distributed.
- MLR assumes little or no multicollinearity (correlation between the independent variable) in data.

Chapter 4

Proposed Methodology

4.1 Aim of Research

We are living in the age of technology and nowadays humans can say that almost anything is possible with the help of technology. Today we have so many tools and methods to access information from any region of this world and Information at this age is so important that without information we would not survive. We have tools that can give us or suggest relevant information at our fingertips and the internet is one of those tools. Today billions of search queries are performed daily and sometimes there given results are relevant and sometimes they are not. In those search queries, thousands of searches are related to medical advice. People often want to know if they have any serious diseases based on their signs and symptoms. But there are no tools available to give them proper information. This research tries to give them tools so that possible disease prediction information can be provided to the end-user.

Some organizations are currently using some sort of decision-making systems but they are not intelligent enough to perform a query to find a disease. They can only perform simple decision-making problems related to the patient's health or disease. Hence this leads to the need for such a decision-making system that can help healthcare organizations to diagnose the exact disease and help the healthcare organizations so that they can provide effective treatment to the patient.

The human body is guarded by the immune system, but sometimes this immune system alone is not capable of preventing our body from diseases. Environmental conditions and living habits of people are the cause of many diseases that are the main reason for a huge number of deaths in the world, and diagnosing these diseases sometimes becomes

challenging. Thus need for a precise, dependable, and feasible system to diagnose such diseases in time for proper treatment is required. With the growth of medical data, many researchers are using these medical data and some machine learning algorithms to help the healthcare communities in the diagnosis of many diseases.

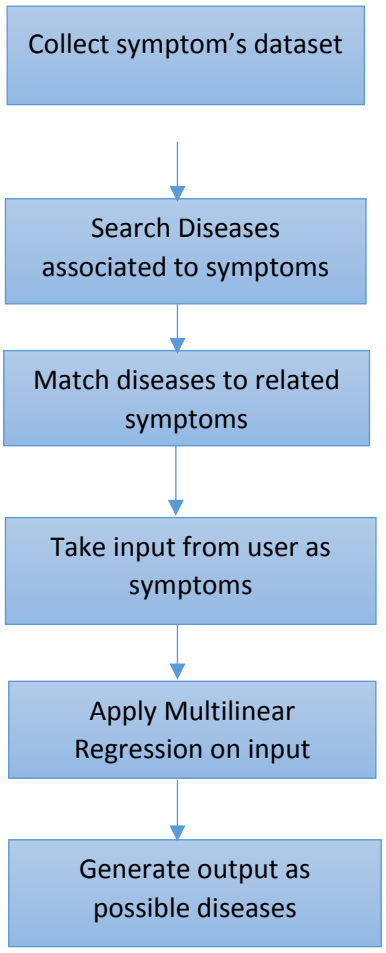
According to McKinsey's [1] report, 50% of Americans have a minimum one or more chronic diseases. Because of the living habits of people, the chances of chronic disease is increasing. In India, as the lifestyle of people has improved, the frequency of diseases is also increased. Approximately 61% of death has occurred due to non-communicable illness like heart disorder, cancer, and diabetes. These diseases are often caused by environmental conditions and living habits of people.

In this research, we are starting an initiative so that disease-related information can be provided to the end-user based on their symptoms. Nowadays we have a humongous amount of data on patient diagnosis and we can use that data to create a disease prediction system that will suggest possible diseases based on patient's symptoms. If a patient has common symptoms that lie within multiple diseases then the system will suggest max possible disease and min possible diseases.

It is a system made using machine learning algorithms to guess the possible diseases based on the patient's symptoms. The growth of information technology has drastically improved our lives. It provides many tools that can save millions of lives, and machine learning is one of them. Machine Learning is used here to develop a system that can help us predict multiple diseases based on symptoms. It can suggest the doctors, probability of the possible diseases. And diagnosis can be done based on suggestion, thus cost could be reduced.

4.2 Proposed Methodology

- First I will collect the datasets of symptoms and their functional problem in the body.
- Then I will collect the information that will associate the symptoms to possible diseases thus related disease information will be collected.
- Then I will get the symptoms as input from the patient and process it by Multilinear Regression.
- After that Multilinear Regression predicts the diseases that may be possible for those acquired symptoms.
- Then the system will show the diagnosis in the form of max possible disease and min possible disease.
- The flow chart of the methodology is given below.



4.3 Algorithm Used

In our disease prediction system, We are using the Support Vector Machine(SVM) for classification and Multilinear Regression(MLR) for predicting the result. MLR is a form of regression algorithm where multiple independent values are involved, meaning that we try to predict a value based on two or more variables.

Simple Linear Regression, where a single Independent/Predictor(X) variable is used to model the response variable (Y). But there may be many situations in which the reply variable is affected by multiple forecaster variables; for such cases, we use the MLR algorithm.

Multilinear Regression is an expansion of Simple Linear regression because it takes multiple forecaster variables to predict the reply variable.

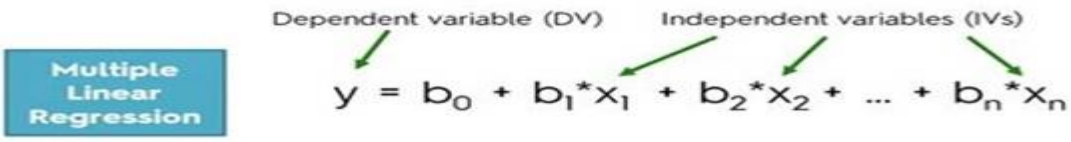
"Multiple Linear Regression is one of the important regression algorithms which models the linear relationship between a single dependent continuous variable and multiple independent variables."

Key facts about Multilinear Regression (MLR):

- MLR, the reliant variable(Y) must be continuous or real, but the forecaster or independent variable may be in the categorical or continuous form.
- Each feature variable must model the linear affiliation with the dependent variable.
- MLR tries to fit a regression line through a multidimensional space of data-points.

MLR equation:

In Multilinear Regression, the target variable(Y) is a linear combination of multiple predictor variables x1, x2, x3, ...,xn. Since it is an enhancement of Simple Linear Regression, so the same is applied for the multiple linear regression equation, the equation becomes:



Where, Y= Output/Response variable

b0, b1, b2, b3 , bn....= Coefficients of the model.

x1, x2, x3, x4,...= Various Independent/feature variable

Assumptions for Multilinear Regression:

- A linear relationship should exist between the Target and predictor variables.
- The regression residuals must be normally distributed.
- MLR assumes little or no multicollinearity (correlation between the independent variable) in data.

4.4 The Architecture of Disease Prediction System

The architecture of DPS includes multilple following fields:

Input: We are taking input from the user of the disease prediction system as a symptoms list.

Get Data

In this field, the user will provide data about their symptoms.

Data Acquisition and Processing

In this field, the input is provided for processing. Data acquisition and processing perform two operations, first is the acquiring the data and then second is the processing of the data and extracting information based on that acquired data.

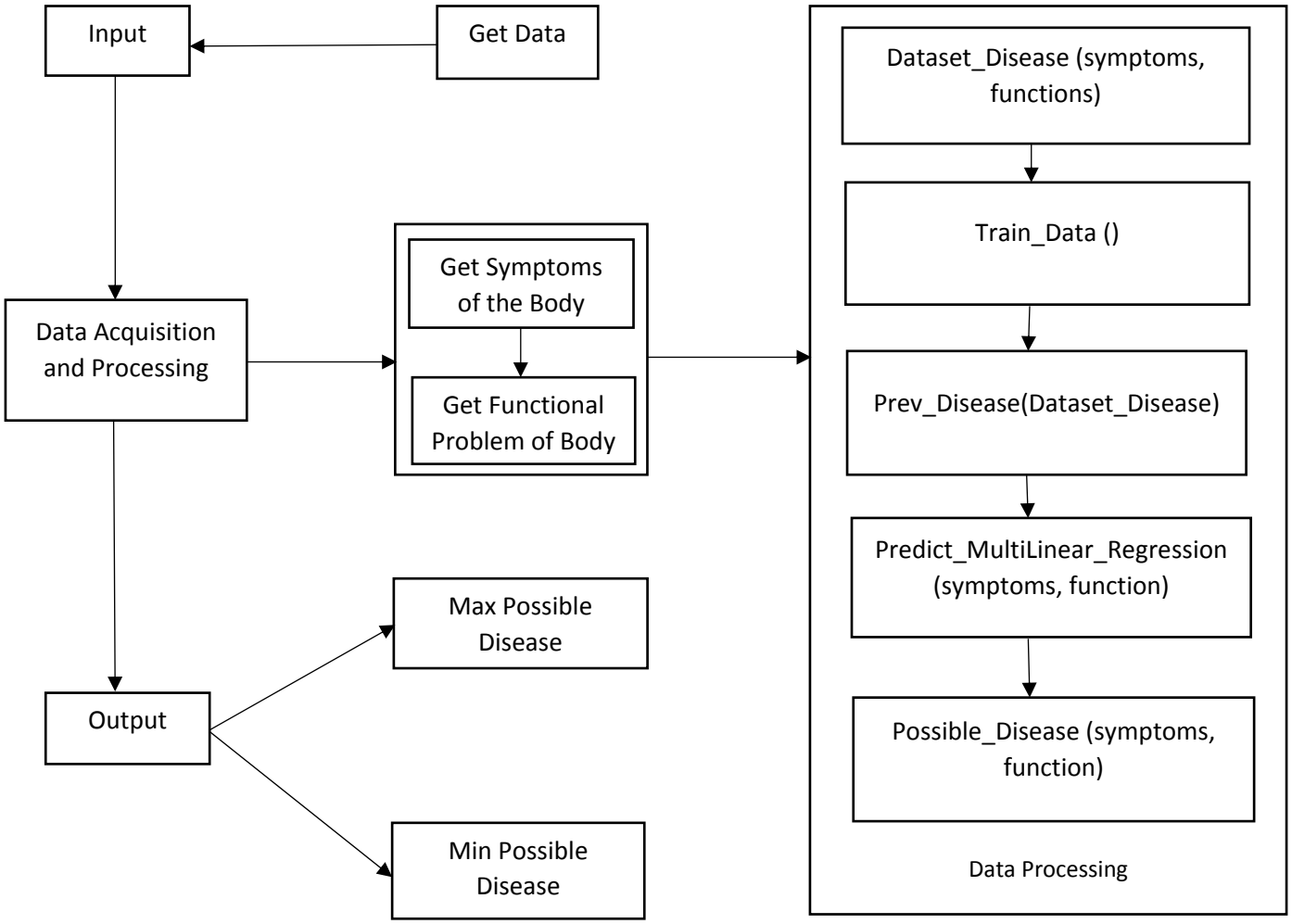


Fig 4.2 The architecture of Disease Prediction System

Get Symptoms of the Body

In this field symptoms of the body are gathered and analyzed. So that this information can be used by the algorithm to predict the possible diseases.

Get Functional Problem of Body

In this field, functional problems of the body that is associated with the symptoms are gathered. So that it is analyzed to get the possible disease.

Dataset_Disease (symptoms, functions)

In this field, we have a predefined dataset of diseases that involves symptoms and functions that are caused by the disease. This dataset is further used to match the data that has been obtained from the user and if matched properly then the system will suggest the possible diseases.

Train_Data ()

In this field training of the system is performed. Our disease prediction system is trained using the SVM (support vector machine) algorithm. Here we are using the SVM algorithm to solve a problem related to regression.

Prev_Disease (Dataset_Disease)

In this field Dataset of the diseases is provided as parameter and processing are performed based on this dataset.

Predict_MultiLinear_Regression (symptoms, function)

In this field, the prediction is performed using the MLR algorithm. In MLR, multiple independent variables are used to perform the prediction of the disease. Symptoms and their functions in the user's body are involved in the prediction.

Possible_Disease (symptoms, function)

In this field symptoms and functions are passed as a parameter and possible diseases are calculated based on these parameters.

Data Processing

This field contains the above five data processing fields and is the main part of our disease prediction system. It has all the necessary fields for processing the data.

Output

After Data Acquisition and Processing, possible diseases are generated as output.

Max Possible Disease

This field contains the maximum possible disease as output.

Min Possible Disease

This field contains the minimum possible diseases as output.

Chapter 5

Result Analysis and Discussion

5.1 ALGORITHM FOR DISEASE PREDICTION SYSTEM

I. Take input of symptoms in p[ ] and their function in t[ ].

II. Declaration:

S[n][m] ← m set of symptoms of n disease

F[n][m] ← m set of functions of n disease

III. $\sum_{i \in I} D(i) \leq \sigma$, and it has a minimum cardinality.

IV. Set S_{svm}[ ] = new set of possible disease symptoms

V. Set F_{svm}[ ] = new set of possible disease functions

VI. For each x, y in S_{svm}[ ] and F_{svm}[ ]

VII. If p[ ] in S_{svm}[ ] and t[ ] in F_{svm}[ ]

VIII. Pdisease[ ] = S_{svm}[ ], F_{svm}[ ], Priority++

IX. endIf

X. endFor

XI. Possible disease Pdisease[0]

Explanation of the above-written algorithm:

- Input is taken from the user in the form of symptom and stored in p[ ] and function is stored in t[ ].
- In the second step, the declaration is done and S[n][m] stores m set of symptoms of n disease and F[n][m] stores m set of functions of n disease.
- $\sum_{i \in I} D(i) \leq \sigma$, it has a minimum cardinality.
- Set S_{svm}[ ] is a new set of possible disease symptoms
- Set F_{svm}[ ] = new set of possible disease functions

- In this step, we are using a for loop to check each value of $S_{svm}[\ ]$ and $F_{svm}[\ ]$
- And if given input lies within $S_{svm}[\ ]$ and $F_{svm}[\ ]$.
- Then we are storing that disease in Pdisease[ ] and increasing the priority of that disease.
- If statement end.
- For loop end.
- The possible disease is given as output Pdisease[0].

5.2 RESULT ANALYSIS

5.2.1 Disease based accuracy analysis for 100 cases

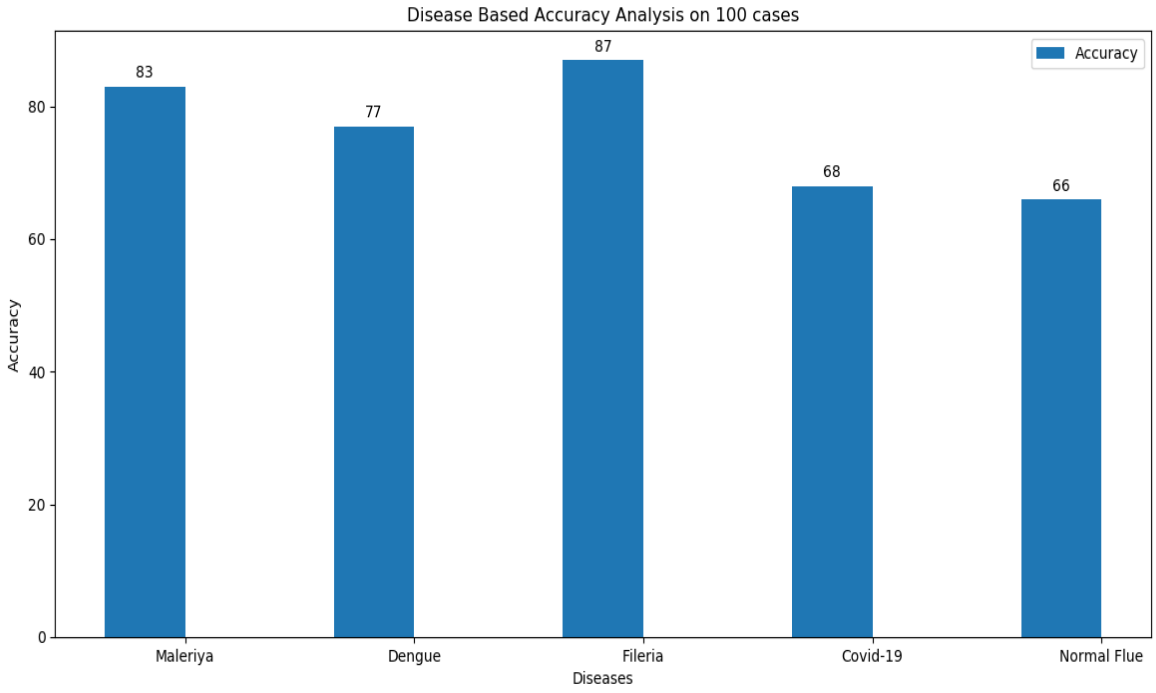


Fig. 5.1. Disease Based Accuracy analysis on 100 cases

Here in this chart, 5 diseases are shown and their accuracy results are shown in the bars. These five diseases are malaria, dengue, filaria, covid-19 and normal flu. And accuracy for malaria is 83% for dengue is 77% and for filaria is 87% and for covid-19 68% and for normal flu is 66%. Accuracy of our system will increase with time when our system will get more and more data then it will be able to easily predict the disease with higher accuracy.

This disease prediction system is trained for 5 diseases in present but our system can be trained for multiple disease prediction. Our algorithm provides higher accuracy for many diseases.

5.2.2 Disease based accuracy analysis for 100 cases using SVM and CNN

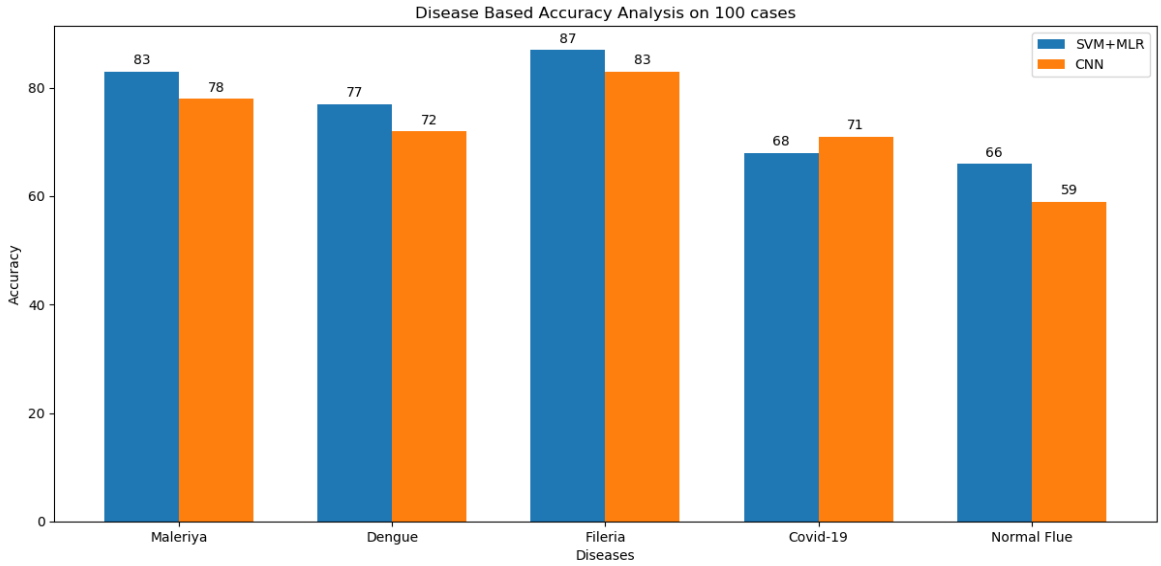


Fig. 5.2. Disease Based Accuracy Analysis on 100 cases comparison

In the above chart we can see that five diseases are given and for these 5 diseases there accuracies are also given. These five diseases are processed using two different algorithms for each consecutive bars. The blue bar shows accuracy for the diseases processed using SVM. The Orange bar shows the accuracy of diseases processed using CNN.

The accuracy for diseases that has been predicted using support vector machine is as follows:

Malaria has 83% accuracy, Dengue has 77% accuracy, Filaria has 87% accuracy, Covid-19 has 68% accuracy and Normal flu has 66% accuracy.

The accuracy of diseases that has been predicted using convolution neural network is as follows:

Malaria has 78% accuracy, Dengue has 72% accuracy, Filaria has 83% accuracy, Covid-19 has 71% accuracy and Normal flu has 59% accuracy.

It is shown in the above chart that both algorithms SVM and CNN are pretty good for suggesting the possible diseases but SVM has a little bit higher accuracy that is why we have used SVM in this research. Support vector machine can be very accurate for many cases but it requires a large amount of data set and if it is provided with huge amount of data sets then it can give more accurate results and it can be used in most of the prediction systems.

5.2.3. Response Time Analysis

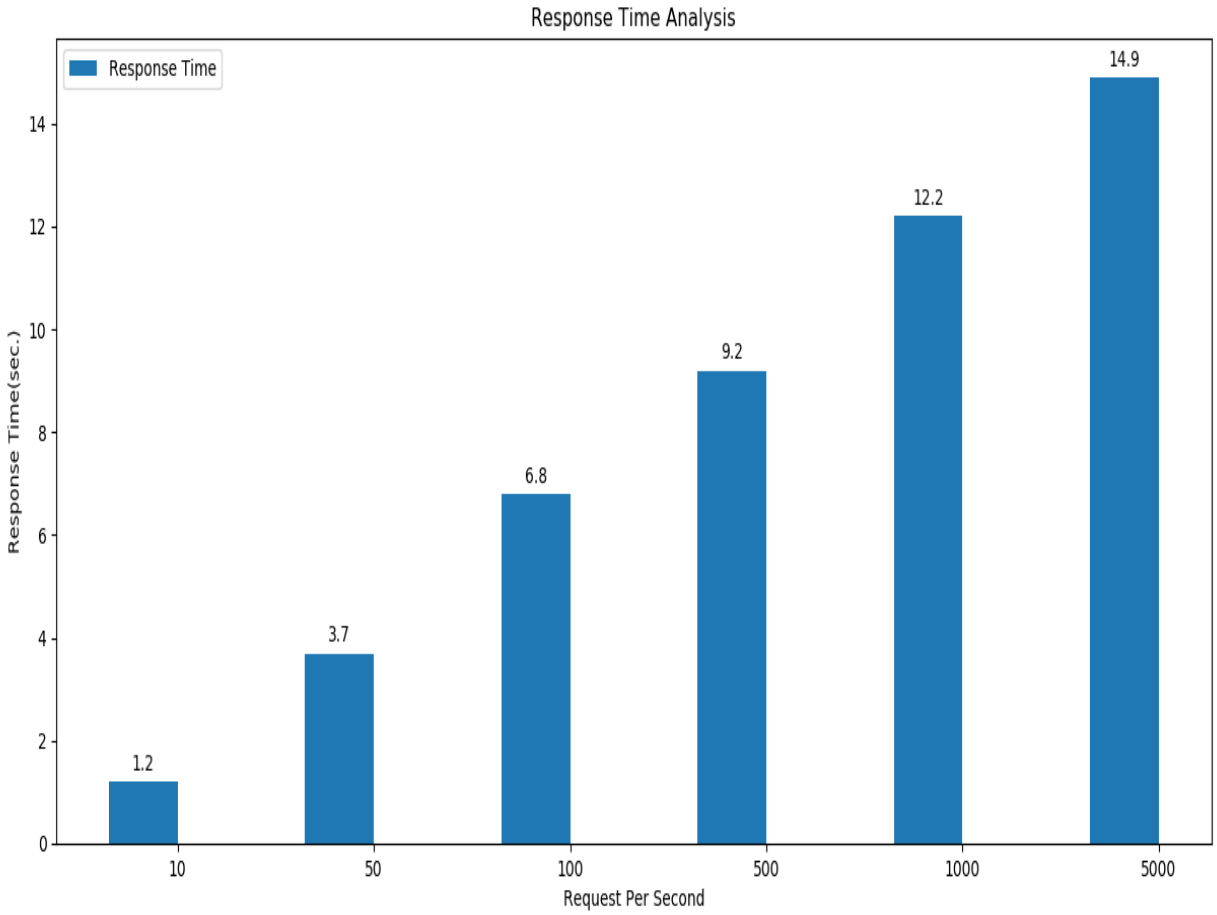


Fig. 5.3. Response Time Analysis

In the above chart x axis shows request per second and y axis shows response time in seconds. Result analysis for response time is very important to show that our system can handle multiple requests at a time. Response time analysis is also important to show that our system is robust enough to handle thousands of request within few seconds and it also shows, how much time system is taking process how many request.

The response time for request are as follows:

For 10 requests our system takes 1.2 second, for 50 requests our system takes 3.7 seconds, for 100 requests our system takes 6.8 seconds, for 500 requests our system takes 9.2 seconds, for 1000 requests our system takes 12.2 seconds, for 5000 requests our system takes 14.9 seconds.

Above response time results shows that this system is fast enough to handle tens of thousands requests within seconds. But for handling this much request our system must be deployed on very powerful cloud computing system so that it can easily process this much request that it is processing locally. Response time can be reduced with higher processing power. And if we will run our system on high end machines then its response time can also be reduced.

5.2.4. Comparative analysis between algorithms for our disease prediction system

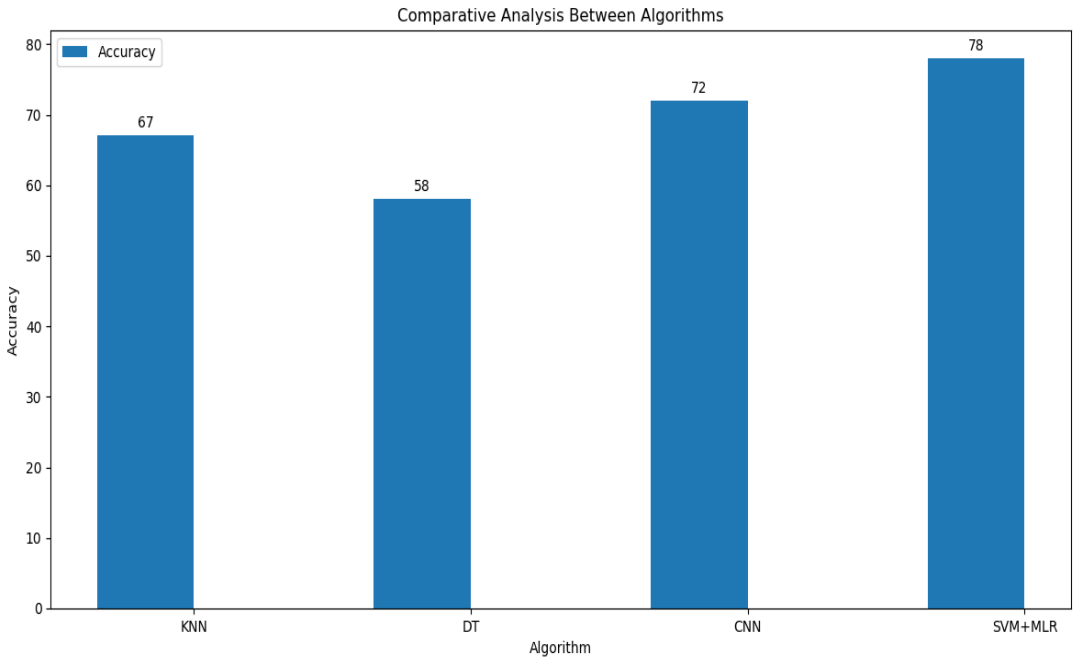


Fig. 5.4. Comparative Analysis between Algorithms

In the above chart, the x-axis shows multiple algorithms and the y-axis shows the accuracy of these algorithms. In the x-axis, we have 4 different algorithms. The first bar shows the accuracy of KNN(K nearest neighbor) algorithm, the second bar shows the accuracy of DT(Decision tree) algorithm, the third bar shows accuracy of CNN(Convolution neural network) algorithm and the last bar shows the accuracy of SVM(Support vector machine) and MLR(Multilinear Regression).

The comparative analysis of different algorithms shows different results for multiple algorithms so that we can compare the result of these algorithms. Based on these comparisons we can decide which algorithm we should use in our systems. A comparative analysis is also important to show that why any research is conducted using the particular algorithm that is used in that research and it can give proof to support why that particular algorithm in that research is used.

The average accuracies of above 4 algorithms are as follows:

K nearest neighbor has average 67% accuracy, decision tree has average 58% accuracy, convolution neural network has average 72% accuracy and support vector machine has average 78% accuracy.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

For this research, I have reviewed multiple research papers and concluded that there should be a system that will be able to predict multiple diseases based on their symptoms it should not be limited to only one or two diseases. Existing systems have limitations that they can only predict one or two diseases and their accuracy is also not much higher. So in my research, I tried to develop a system that will be able to predict many diseases and I have taken 5 diseases examples in my system and their accuracies can go up to 87%.

With the help of a disease prediction system, it is easy to get a possible disease diagnosis using this system. It provides easy access for normal people and fresher doctors to know the possible diseases based on symptoms. It can suggest to users several possible diseases and which disease has a higher possibility for a set of symptoms. It can reduce the cost of diagnosing multiple diseases and based on the possibility of the disease provided by the disease prediction system doctors can get patients diagnosed for the minimum number of diseases, thus cost is reduced.

With the help of a disease prediction system, it was possible to diagnose people based on symptoms. Disease prediction system provides only possible outcomes it does not guarantee that it will predict the disease correctly. But it has significantly higher accuracy for predicting possible diseases. In our research, we have analyzed the accuracy of this system for 5 different diseases and our accuracy can go up to 87%.

In our research, we have used a support vector machine algorithm and multilinear regression algorithm to predict diseases. And we have also tested multiple algorithms like the k-nearest

neighbor, convolution neural network, decision tree, etc. Despite testing these algorithms I have found that the support vector machine gives higher accuracy than other algorithms.

The purpose of this research was to provide medical diagnosis information to normal people, fresher doctors, medical students, and anyone who wants to know about a set of symptoms and associated diseases.

In this research, we have found that possible disease prediction can go up to 87% for some diseases and minimum 68% for some diseases but these results are obtained using the minimum amount of data set but if we can feed the system humongous amount of data set then this disease prediction system can give accuracy up to 95%. Obtaining a humongous amount of data set related to diseases and their symptoms is very time consuming and it cannot be done within one or two years it requires multiple years to collect those data sets and train the system using those data searches. This system can be used by Ph.D. scholars to do further research.

This system can run on very small computing power but if we want to handle multiple requests at the same time then it's processing/computing power should we little bit higher. While deploying our system on the internet we have to keep it in mind that our system must be deployed in the cloud so that it can receive high computing power and give results faster because cloud computing provides multiple facilities and one of them is processing power and handling thousands of request at a time requires a high amount of computing and processing power that can only be obtained from a dedicated server or by using the cloud, so we will choose the cloud for our system because it is very cost-effective and we have no headache to maintain it.

6.2. Future Scope

This system has a very broad future scope because it can be used by multiple medical organizations and research institutions to commercialize disease prediction system in their existing systems and provide a relevant diagnosis to the patient and reduce the cost for multiple disease tests.

In this system, we have only used 5 diseases to show the functionality of this system but in the future, so many diseases can be added to the system and their diagnosis would be possible. This system is very fast for suggesting the possible diseases just based on the symptoms. Medical students can also use this system to learn which symptoms cause which disease.

For our system to be accurate it needs to be trained using multiple data sets and these data sets should be in large quantity. Once our system is trained for a disease then it would be able to predict that particular disease based on their symptoms and to predict many diseases we have to train our system with multiple disease data sets in very large quantities. Our diseases list can go to any length because this system has the capability to be trained for any number of diseases it just has one limitation that is the number of data sets if we provide a large number of data sets to predict the disease then its accuracy will be higher but if you provide less number of data sets then its accuracy will be lesser.

REFERENCES

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities" IEEE Access, vol. 5, no. 1, pp. 8869–8879, 2017.

[2] Sayali Ambekar, Rashmi Phalnikar, "Disease Risk Prediction by Using Convolutional Neural Network" IEEE, 978-1-5386-5257-2/18, 2018.

[3] Naganna Chetty, Kunwar Singh Vaisla and Nagamma Patil, "An Improved Method for Disease Prediction using Fuzzy Approach" IEEE, DOI 10.1109/ICACCE.2015.67, pp. 569-572, 2015.

[4] Dhiraj Dahiwade, Gajanan Patle and Ektaa Meshram, "Designing Disease Prediction Model Using Machine Learning Approach" IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4, pp. 1211-1215, 2019.

[5] Lambodar Jena and Ramakrushna Swain, "Chronic Disease Risk Prediction using Distributed Machine Learning Classifiers" IEEE, 978-1-5386-2924-6/17, pp. 170-173, 2017.

[6] Dhomse Kanchan B. and Mahale Kishor M., "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis" IEEE, 978-1-5090-0467-6/16, pp. 5-10, 2016.

[7] Pahulpreet Singh Kohli and Shriya Arora, "Application of Machine Learning in Disease Prediction" IEEE, 978-1-5386-6947-1/18, pp. 1-4, 2018.

[8] Deeraj Shetty, Kishor Rit, Sohail Shaikh and Nikita Patil, ” Diabetes Disease Prediction Using Data Mining” IEEE, 978-1-5090-3294-5/17, 2017.

[9] Rashmi G Saboji and Prem Kumar Ramesh,“A Scalable Solution for Heart Disease Prediction using Classification Mining Technique” IEEE, 978-1-5386-1887-5/17, pp. 1780-1785, 2017.

[10] Rati Shukla, Vikash Yadav, Parashu Ram Pal and Pankaj Pathak, "Machine Learning Techniques for Detecting and Predicting Breast Cancer" IJITEE, ISSN: 2278-3075, Volume-8, pp. 2658-2662, 2019.

[11] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” IEEE Access, DOI 10.1109/ACCESS.2019.2923707, pp. 81542-81554, 2019.

[12] Anjan Nikhil Repaka, Sai Deepak Ravikanti and Ramya G Franklin,”Design And Implementing Heart Disease Prediction Using Naives Bayesian” IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8, pp. 292-297, 2019.

[13] Aakash Chauhan, Purushottam Sharma, Vikas Deep and Aditya Jain, “Heart Disease Prediction using Evolutionary Rule Learning” CICT 2018.

[14] Aditi Gavhane, Gouthami Kokkula, Isha Pandya and Kailas Devadkar, “Prediction of Heart Disease Using Machine Learning” IEEE Xplore ISBN: 978-1-5386-0965-1, pp. 1275-1278, 2018.

[15] Ankita Dewan and Meghna Sharma, “Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification” IEEE, 978-9-3805-4416-8/15, pp. 704-706, 2015.

[16] PU, J. and LEI, X., 2013. Two-stage fast training method based on core vector machine and support vector machine. Journal of Computer Applications, 32(2), pp.419-424.

[17] International Journal of Science and Research (IJSR), 2016. Breast Cancer Classification using Support Vector Machine and Neural Network. 5(3), pp.1-6.

[18] International Journal of Science and Research (IJSR), 2016. Breast Cancer Classification using Support Vector Machine and Neural Network. 5(3), pp.1-6.

[19] IEEE Cloud Computing, 2016. IEEE Cloud Computing Call for Papers. 3(5), pp.63-63.

[20] Mathews, A., 2019. What can machine learning do for information security?. Network Security, 2019(4), pp.15-17.

[21] BMJ, 1893. The Cholera Epidemic: "Mixed Infection" in Cholera. 1(1676), pp.317-318.

[22] Bhamidimarri, K., Park, J. and Dieterich, D., 2011. Management of Hepatitis B Virus Coinfection: HIV, Hepatitis C Virus, Hepatitis D Virus. Current Hepatitis Reports, 10(4), pp.262-268.

[23] BUNYAN, I., 2020. Corona Virus Affected of Human and Animal: A Review. Journal of Research on the Lepidoptera, 51(2), pp.1116-1134.

[24] Computers & Security, 1984. Data protection-data security-privacy. 3(1), pp.57-58.

[25] Sergent, M., Mathieu, D., Phan-Tan-Luu, R. and Drava, G., 1995. Correct and incorrect use of multilinear regression. Chemometrics and Intelligent Laboratory Systems, 27(2), pp.153-162.

Plagiarism Check Report

Chapter 1

PLAGIARISM SCAN REPORT

Words	769	Date	August 09,2020
Characters	4959	Exclude Url	

11% Plagiarism	89% Unique	4 Plagiarized Sentences	33 Unique Sentences

PLAGIARISM SCAN REPORT

Words	800	Date	August 09,2020
Characters	5111	Exclude Url	

13% Plagiarism	87% Unique	5 Plagiarized Sentences	34 Unique Sentences

Chapter 2

PLAGIARISM SCAN REPORT

Words	630	Date	August 08,2020
Characters	3624	Exclude Url	

0% Plagiarism	100% Unique	0 Plagiarized Sentences	34 Unique Sentences

SmallSEOTools

PLAGIARISM SCAN REPORT

Words	984	Date	August 08,2020
Characters	6853	Exclude Url	

0%	100%	0	43
Plagiarism	Unique	Plagiarized Sentences	Unique Sentences

SmallSEOTools

PLAGIARISM SCAN REPORT

Words	946	Date	August 08,2020
Characters	6357	Exclude Url	

0%	100%	0	43
Plagiarism	Unique	Plagiarized Sentences	Unique Sentences

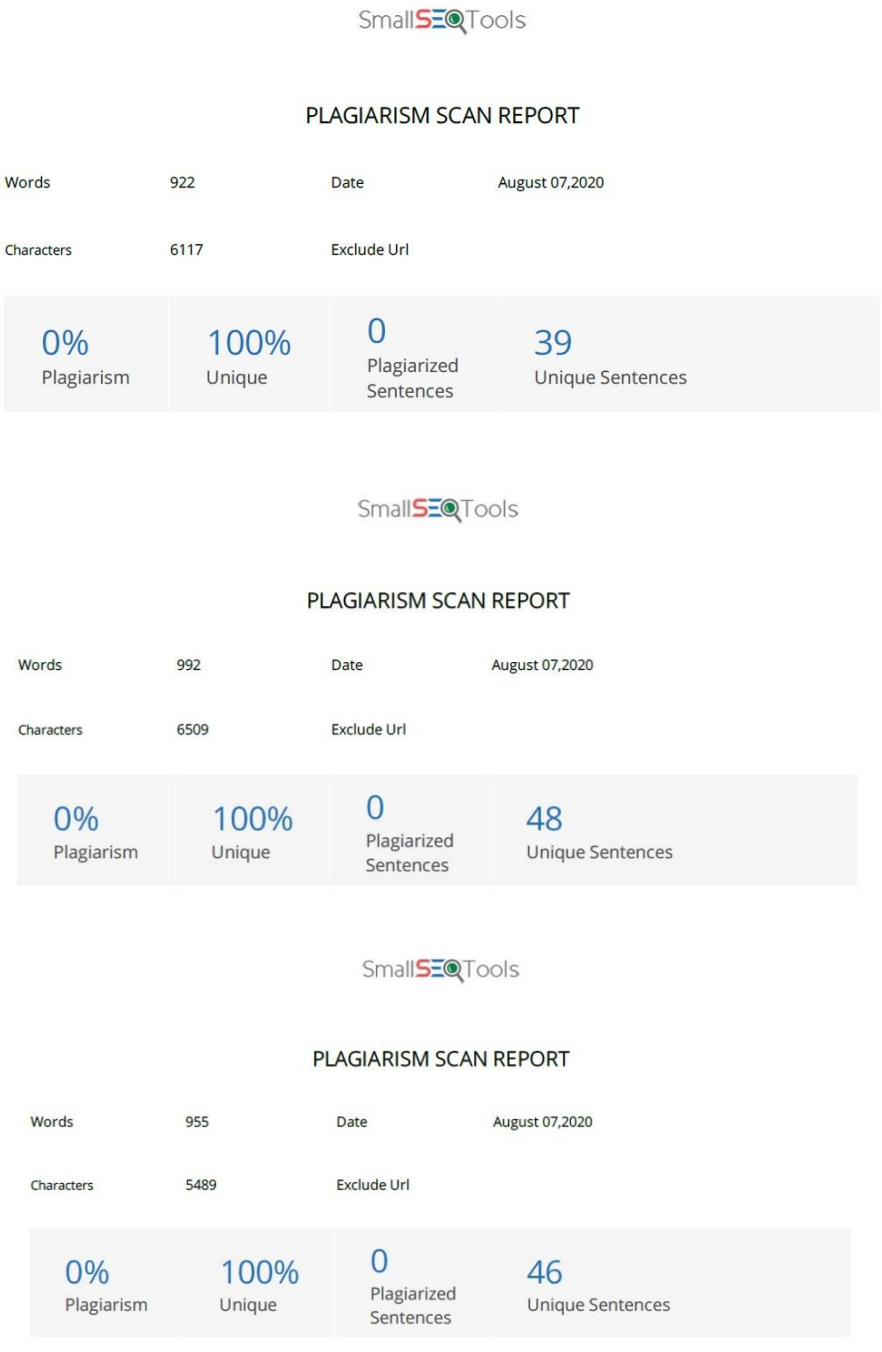
SmallSEOTools

PLAGIARISM SCAN REPORT

Words	929	Date	August 08,2020
Characters	5974	Exclude Url	

0%	100%	0	42
Plagiarism	Unique	Plagiarized Sentences	Unique Sentences

Chapter 3



SmallSEOTools

PLAGIARISM SCAN REPORT

Words	922	Date	August 07,2020
Characters	6117	Exclude Url	

0% Plagiarism	100% Unique	0 Plagiarized Sentences	39 Unique Sentences

SmallSEOTools

PLAGIARISM SCAN REPORT

Words	992	Date	August 07,2020
Characters	6509	Exclude Url	

0% Plagiarism	100% Unique	0 Plagiarized Sentences	48 Unique Sentences

SmallSEOTools

PLAGIARISM SCAN REPORT

Words	955	Date	August 07,2020
Characters	5489	Exclude Url	

0% Plagiarism	100% Unique	0 Plagiarized Sentences	46 Unique Sentences


SmallSEOTools
PLAGIARISM SCAN REPORT
Words 955
Date August 07,2020
Characters 5489
Exclude Url
0% Plagiarism
100% Unique
0 Plagiarized Sentences
46 Unique Sentences

Chapter 4


SmallSEOTools
PLAGIARISM SCAN REPORT
Words 944
Date August 07,2020
Characters 5992
Exclude Url
0% Plagiarism
100% Unique
0 Plagiarized Sentences
47 Unique Sentences

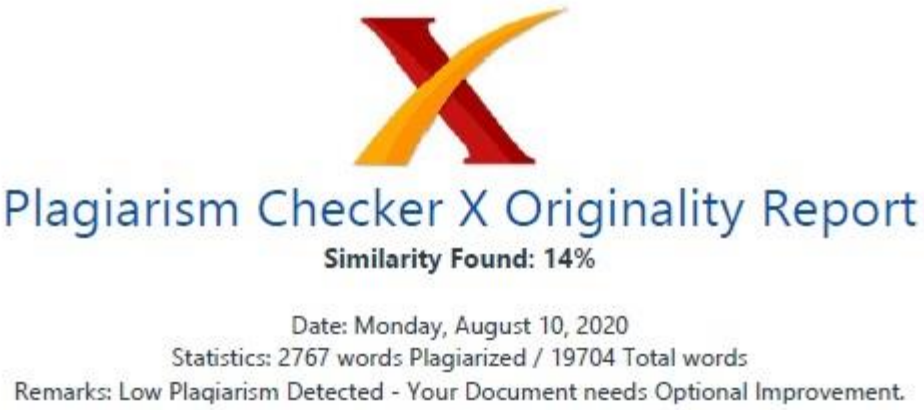

SmallSEOTools
PLAGIARISM SCAN REPORT
Words 422
Date August 07,2020
Characters 2732
Exclude Url
0% Plagiarism
100% Unique
0 Plagiarized Sentences
25 Unique Sentences

Chapter 5

Chapter 6

Plagiarism Check Report of Plagiarism Checker X

All chapters



Publication from This Work

1. "**A Detailed Review on Disease Prediction Models that uses Machine Learning**" has been published in International Journal of Innovative Research in Computer Science & Technology (IJIRCST). Certificate of publishing is below.


DOI: 10.21276/ijircst ISSN: 2347-5552

International Journal of Innovative Research in Computer Science & Technology
(A Peer-Reviewed, Referred, & CrossRef Indexed Journal)
www.ijircst.org

Certificate of Publishing

This is to certify that
Md. Ehtisham Farooqui
has contributed as an author of paper entitled
"A Detailed Review on Disease Prediction Models that uses Machine Learning"
Published in Volume-8, Issue-4, July-2020

2. "**Disease Prediction System using Support Vector Machine and Multilinear Regression**" has been published in International Journal of Innovative Research in Computer Science & Technology (IJIRCST). Certificate of publishing is below.


DOI: 10.21276/ijircst ISSN: 2347-5552

International Journal of Innovative Research in Computer Science & Technology
(A Peer-Reviewed, Referred, & CrossRef Indexed Journal)
www.ijircst.org

Certificate of Publishing

This is to certify that
Md. Ehtisham Farooqui
has contributed as an author of paper entitled
"Disease Prediction System using Support Vector Machine and Multilinear Regression"
Published in Volume-8, Issue-4, July-2020