Bachelor of Science in Computer Science & Engineering

A Machine Learning Approach to Clinically Diagnose Human Pyrexia Cases

by

Dipon Talukder

ID: 1504009

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

April, 2021

A Machine Learning Approach to Clinically Diagnose Human Pyrexia Cases

Submitted in partial fulfilment of the requirements for
Degree of Bachelor of Science
in Computer Science & Engineering

by

Dipon Talukder

ID: 1504009

Supervised by

Dr. Md. Mokammel Haque

Professor

Department of Computer Science & Engineering

Chittagong University of Engineering & Technology (CUET)

Chattogram-4349, Bangladesh.

The thesis titled '**A Machine Learning Approach to Clinically Diagnose Human Pyrexia Cases'** submitted by ID: 1504009, Session 2019-2020 has been accepted as satisfactory in fulfilment of the requirement for the degree of Bachelor of Science in Computer Science & Engineering to be awarded by the Chittagong University of Engineering & Technology (CUET).

Board of Examiners

Chairman

Dr. Md. Mokammel Haque
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (Ex-Officio)

Dr. Md. Mokammel Haque
Professor & Head
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Member (External)

Dr. Asaduzzaman
Professor
Department of Computer Science & Engineering
Chittagong University of Engineering & Technology (CUET)

Declaration of Originality

This is to certify that I am the sole author of this thesis and that neither any part of this thesis nor the whole of the thesis has been submitted for a degree to any other institution.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. I am also aware that if any infringement of anyone's copyright is found, whether intentional or otherwise, I may be subject to legal and disciplinary action determined by Dept. of CSE, CUET.

I hereby assign every rights in the copyright of this thesis work to Dept. of CSE, CUET, who shall be the owner of the copyright of this work and any reproduction or use in any form or by any means whatsoever is prohibited without the consent of Dept. of CSE, CUET.

Signature of the candidate

Date:

Acknowledgements

The study acknowledges and extends tremendous appreciation to all individuals without whom the study would not be possible to finish. First and foremost, I would like to express my gratitude to my supervisor, Professor Md. Mokammel Haque, Ph.D., Department of CSE, Chittagong University of Engineering and Technology, for his unwavering encouragement and confidence in my work during the study.

Also, I cannot express enough thanks to the persons without whom the study could not be accomplished, Dr. Puspen Das Gupta, Dr. Rubiath Farhin Etu, Dr. Diptanil Das, Dr. Joy Deb, Dr. Mejbah Uddin, Dr. Md. Hossain Al Imran and many more who have helped to develop the dataset by collecting symptoms from the patients and for expert support.

Abstract

For the past decade, a lot of research has been conducted to employ the ability of machine learning in the healthcare system to diagnose patients with diseases. It has been observed that incorporating machine learning into a modern diagnostic method will significantly reduce the chance of misdiagnosing outcomes. To have an impact on the healthcare system, this study provides a technique for diagnosing the three most closely symptomized diseases: Dengue, Malaria, and Typhoid. In order to develop the diagnosing model, symptoms of pyrexia patients admitted to the hospital have been obtained from five districts of Bangladesh. In addition, the study suggests a collection of 29 features for correctly classifying pyrexia cases. The collected data is analyzed, and the key findings of the study are discussed. The dataset is used to test a variety of machine learning algorithms. Despite the fact that many studies have found Support Vector Machine and Rough Set Theory to be the most efficient models in such classification tasks, the work found and recommends linear regression to be the most effective, with a classification accuracy of 95.1%.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence. 6

AROW Adaptive Regularization of Weights. 7

CSS Cascading Style Sheet. 11

CWL Confidence Weighted Learning. 7

HTML HyperText Markup Language. 10

ML Machine Learning. 6, 7

RL Reinforcement Learning. 36

RST Rough Set Theory. 6, 29

SVM Support Vector Machine. 7, 8, 29

WSGI Web Server Gateway Interface. 12

Chapter 1

Introduction

1.1 Introduction

Pyrexia is a condition where body temperature accelerates in an abnormal manner. Pyrexia is also commonly known as fever. Mild escalation of body temperature can be resolved by in-taking counter medication available such as Ibuprofen and Paracetamol. But a sudden high rise in body temperature can be life-threatening. Although there is variance among expertise about the right body temperature for humans but the value ranges from 99.0 to 100.9 degrees Fahrenheit [1, 2, 3]. Pyrexia is considered to be one of the most regular medical signs. It occurs in up to 75 percent of chronically ill adults [4]. Other than adults, pyrexia accounts for about 30 percent of children's visits to the doctor [5].

Among many explanations Typhoid, Dengue, Malaria are regarded as common causes of pyrexia [6]. 12.5 million people were infected globally in 2015 alone by Typhoid [7]. The same year, 1.49M deaths are reported worldwide due to Typhoid [8]. Without treatment, the probability of mortality may be as high as 20% in this disease [9]. On the other hand, in the case of dengue, around 390 million new cases are registered every year among which approximately 40 thousand people die [10, 11]. Malaria caused 228 million cases worldwide in 2018, with an approximate 405,000 deaths [12]. These infectious diseases are often misdiagnosed [13].

Between numerous applications, machine learning technology is showing promising results in the development of the healthcare system for the past few years [14, 15, 16]. Machine Learning has the ability to find a hidden pattern among the complex medical data that is creating an impact on clinical diagnosis systems.

This chapter overviews the proposed system to diagnose human pyrexia cases and challenges faced upon completion of the work. The chapter also summarizes the motivation and contribution of the work to the healthcare system.

1.2 Work Overview

Machine Learning Systems require data to feed into the machine learning model. Traditional machine learning models are sensitive to feature selection and the way data is preprocessed. Hence, before data collection, feature selection is an important step. Symptoms of pyrexia patients are recorded according to the feature. The collected string data is mapped to its equivalent numerical data for computation purposes. The numerical data is preprocessed and visualized to understand the nature of the data.

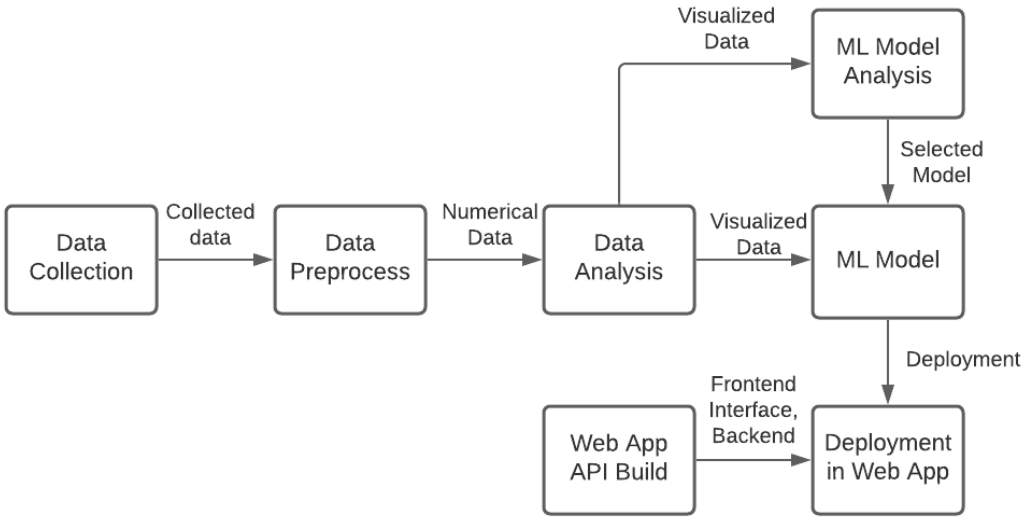


Figure 1.1: System Overview

The data is then fed into numerous machine learning models to analyze and select a machine learning model with the highest accuracy for the better performance of the system. A Web App is developed using frontend and backend web technologies. The selected machine learning model is deployed to the web app with the ambition of high usability.

1.3 Challenges

The task of diagnosing human pyrexia cases developed several obstacles that have to be overcome.

Firstly, Differentiating the three most closely symptomized infectious diseases requires prudent design and careful choice of features [17]. Obtaining accurate learning outcomes is a critical step in mission definition. Outcomes are often used to generate the gold-standard labels used for supervised prediction activities.

Secondly, the Machine Learning model provides results based on the data experienced itself [18]. But sensitive issues like diagnosis need data recorded from patients admitted to the hospital. Again, Data has to be collected according to the feature selected.

Thirdly, Mislabeled data can greatly degrade the performance of a machine learning model. Models try to set a relationship between the features selected and the outcome result. Information leakage can make prediction meaningless.

1.4 Applications

There are several areas in the healthcare system where machine learning technology is vastly being used for its property of finding relation through the unstructured statistical data. The impact of this technology in the clinical system is likely to increase in the future. The technology is evolving and as more data is generated and recorded this can considerably reduce the hurdle faced presently in the system. Application of the work can be quantified in the following manner:

- Diagnostic aid used to help doctors in real-time by retrieving data or recommending diagnoses.
- Fast diagnostic results can be obtained in areas where diagnostic facilities are not available.
- A core part of the healthcare system is the treatment process. The treatment process can be developed using the proposed system.

1.5 Motivation

Machine Learning and Artificial Intelligence has an immense impact on modern clinical system. The dependency of healthcare systems on technology like machine learning is increasing day by day. According to experts, AI applications that are contributing to the health care can be divided into three aspects- Patient oriented, Clinician oriented, Operational oriented [19]. Motivation of our clinician oriented work can be listed as follows:

1. Undiagnosed or Misdiagnosed cases in rural area
2. Low productivity and low quality of clinical care
3. Lack of proper treatment in time
4. Costly and less time effective solution

1.6 Contribution of the thesis

The work is accomplished to achieve a specific set of goals. The main objective of this work is to contribute and create an impact in the healthcare system through the power of machine learning and web technology. Also, contribution to the machine learning society for the further development of the clinical system. The contributions can be summarized as follows:

1. Developed a dataset containing symptoms of pyrexia patients collected from patients admitted at hospitals, all over Bangladesh.
2. Proposing a set of features (29 Features) to differentiate three closely symptomized infectious diseases (Dengue, Malaria, Typhoid).
3. Analysis of data collected from hospitals
4. Machine Learning model performance comparison on the collected data and recommending the appropriate model for the task.
5. Deployment of the model through Web Technology

1.7 Thesis Organization

The rest of the report is sequenced as follows:

In the next chapter, the researchers' study into applying machine learning in the healthcare sector for disease diagnosis is addressed. The experimental techniques, data collection process, flaws, and algorithms used in the literature for the results are listed.

In chapter 3, the approach used to accomplish the work's goal is well explained and debated. The method begins with the collection of primary features.

In chapter 4, we've clarified how our methods yielded the results we did. The dataset is explained at the beginning of the chapter. Also, the results of the data collection and main conclusions from the data are discussed.

In chapter 5, we concluded the whole work and briefly summarized the impact of our work future works that can be done.

1.8 Conclusion

In this chapter, a high-level overview of the work has been presented. The chapter introduced related concepts to the reader. Also, the challenge faced and the application of this work are briefly discussed. The importance of the work is mentioned in the motivation section and contribution of the work is also provided. In the next chapter, the works of the previous researchers in this area of research are explored.

Chapter 2

Literature Review

2.1 Introduction

In the previous chapter, a summary and the importance of the work are presented. Furthermore, the contribution and application of the work are also discussed. For the past decade, a lot of researches has been conducted for applying ML and AI in the healthcare sector especially for diagnosing disease among patients to improve the efficiency of the existing system. In this chapter, relevant researches of the researchers are evaluated and the structure is outlined. Moreover, the key findings of the researches and the methods that are applied to extract the result are also discussed. Outcomes of the scholarly articles are analyzed to understand the current state.

2.2 Related Healthcare Researches

Practitioners can detect diseases by looking at physiological evidence, environmental conditions, and genetic factors. Machine learning enables the development of models that link a wide variety of features to a disease. In this literature, Oguntimilehin et al. [20] diagnosed five classes of typhoid using a machine learning approach. Data used to train the model has been collected from Typhoid patients in Nigeria. Symptoms are collected and labeled by medical experts. In the literature, authors have quoted medical experts that suggested, although typhoid shares the same symptoms as malaria and dengue still they are differentiable with a large number of feature variables. The researchers used a rule-based RST algorithm and defined 18 rules to classify the level of Typhoid. In the literature, it is shown that researchers were able to achieve 96% accuracy in the test

dataset. But the work holds a huge drawback. Symptoms of the patients are collected from only one hospital hence the dataset contains the typhoid symptoms of only a small region which caused a lack of variance in the dataset.

In another literature, Davi et al [21] diagnosed the severity of dengue fever using a machine learning approach. In this case, researchers used SVM algorithm to locate the optimal loci classification subset and used a Neural Network for the classification of acute dengue fever or moderate dengue fever. For this purpose, the authors used human genome data instead of symptomatically data of the dengue patients. The work also claimed that the proposed methodology can be applied to diagnose any genetically influenced disease.

Srivastava et al. [22] proposed a novel method for classifying dengue fever using an online learning method. The proposed system is initially trained the model with a very few number of training samples but it is able to gain experience with time through the online learning method. Initial training, which is referred to in the literature as offline training is done using SVM and the Random Forest algorithm. The online training is done using AROW and CWL algorithm. The researchers achieved 98% accuracy for predicting dengue patients. 13 features are selected for ML model. However, the system requires the model to integrate with a healthcare system software for collecting symptoms of the patients for the online learning process. Also, no validation is done before labeling the data instances before training.

Fatima et al. [23] surveyed various machine learning algorithms to diagnose various diseases. The experimenters analyzed and recommended appropriate algorithm for particular disease detection. The analysis showed Rough Set Theory algorithm to be most effective for predicting infectious diseases such as dengue.

Chen et al. [24] implemented a machine learning approach to classify acute and Non-acute Covid 19 patients. The scientists have carefully chosen 26 features from the symptom list of covid patients. Symptoms of the patients are recorded from Wuhan, the location where the disease has originated. Researchers found Random Forest Algorithm provided the best accuracy in this regard and pulled of 90% accuracy in classification results.

Lee et al. [25] developed a model to diagnose malaria using patient information. The researchers have compared the developed dataset using six different machine learning models. According to their study, the random forest has proved to be the best model in this regard. The authors also concluded that Machine Learning technology can be successfully applied to predict malaria using the information gathered from the patient.

Oguntimilehin et al. [26] recommended a machine learning approach to predict and provide treatment of Typhoid cases. The authors implemented a decision tree algorithm for the task and achieved 95% accuracy over the test dataset.

In another research paper, Sajana et al [27] proposed a hybrid system to classify disease caused by infection. The authors originally classified two infectious diseases- Dengue, Typhoid. The collected dataset has been fed through many traditional machine learning algorithms to compare the result. 97.2% accuracy has been achieved through their work.

2.3 Machine Learning Algorithms

2.3.1 Support Vector Machine

Support vector machines are supervised machine learning algorithms that can be used for classification and regression tasks [28, 29]. They are mostly used for classification problems. A support vector algorithm plots each acquired data instance as a point on an n-dimensional space or graph, where 'n' is the total number of data features that are present. Each data point's value is expressed by a specific coordinate on the graph. The SVM generates an n-1 dimensional hyperplane to separate the data instances. The technique that is followed in the algorithm is called kernel trick which transforms data upon which an optimal boundary is depicted to separate the instances.

2.3.2 Decision Tree

Decision Tree is one of the most popular supervised machine learning algorithms used to solve a classification problem or a regression problem. This algorithm

aims to construct a model that predicts the value of a target variable, and the decision tree solves the problem by using the tree representation, where the leaf node corresponds to a class label and attributes are expressed on the internal node of the tree. A tree is constructed by dividing the source set, which is the tree's root node, into subsets, which are the tree's successor children. The splitting is built on a series of classification feature-based splitting laws. Because of their comprehensibility and simplicity, decision trees are one of the most common machine learning algorithms [30].

2.3.3 Naïve Bayes

Naive Bayes is a concise method for building classifiers, which are models that assign class labels to problem instances represented as vectors of feature values, with the class labels drawn from a finite set [31]. The algorithm is based on the Bayes probability theorem widely used for classification tasks and statistical problems. For training such classifiers, there is no single algorithm, but rather a family of algorithms based on the same principle: all naive Bayes classifiers presume that the value of one function is independent of the value of every other feature, provided the class variable.

2.3.4 K-Nearest Neighbors

K-Nearest Neighbors is one of Machine Learning's most simple yet crucial classification algorithm [32]. Pattern recognition, data processing, and intrusion detection are only a few of the applications it sees in the supervised learning domain. It is commonly used in real-world situations because it is non-parametric, which means it has no inherent assumptions regarding data delivery. Prior data are also known as training data is provided, which divides coordinates into categories based on an attribute. Initially, it chooses n number of centroids where n is the number of class labels to be predicted. Data points which are closer to one centroid are considered as under the same class. Each data point belongs to one of the clusters that is assigned. The distance is calculated using the Euclidean distance formula.

2.3.5 Logistic Regression

Logistic regression is another statistical methodology that machine learning has adopted [33]. Under the Supervised Learning technique, one of the most common Machine Learning algorithms is logistic regression. It's a method for estimating a categorical dependent variable from several independent variables. Since it can have probability and identify new data using both continuous and discrete datasets, logistic regression is a powerful machine learning algorithm. Except for how they are used, Logistic Regression is somewhat similar to Linear Regression. To solve regression problems, Linear Regression is used, while Logistic Regression is used for solving classification problems.

2.3.6 Random Forest

Random Forest is a well-known machine learning algorithm that uses the supervised learning method [34]. In machine learning, it can be used for both classification and regression problems. It is based on ensemble learning, which is a method of integrating several classifiers to solve a complex problem and increase the model's accuracy. The name forest is as such because the algorithm generates multiple decision trees for the prediction task. Instead of depending on a single decision tree, the random forest takes the predictions from each tree and forecasts the final performance based on the majority votes of predictions. The greater the number of trees in the forest, the more accurate it is and the issue of overfitting is avoided.

2.4 Web Technology

2.4.1 Frontend Technology

- **HTML**

  HTML, or HyperText Markup Language, is the basic markup language for documents that are intended to be used in a web browser. Online browsers

receive HTML documents from a web server or locally stored files and convert them to multimedia web pages. HTML originally provided clues for the document's presentation and defined the layout of a web page semantically. HTML elements are the components that make up HTML pages. Photos and other artifacts, such as interactive forms, may be inserted in the rendered page using HTML constructs. By denoting structural semantics for text such as headings, paragraphs, links, quotations, and other objects, HTML allows the construction of organized documents. The HTML tags are not seen by browsers, but they are used to interpret the page's text.

- **CSS**

  CSS is a style sheet language for defining the appearance of a text written in a markup language like HTML. Along with HTML and JavaScript, CSS is a key component of the World Wide Web. CSS is a style sheet that allows you to separate presentation from text, including layout, colors, and fonts. This separation helps improve content accessibility, provide more flexibility and control in the specification of presentation characteristics, and enable several web pages to share formatting by specifying the relevant CSS in a separate.css file, which reduces complexity and repetition in the structural content and allows the .css file to be cached to improve page load speed between the pages that have the same format and similar file.

2.4.2 Backend Technology

- **Flask**

  Flask is a Python-based microweb framework. It is referred to as a microframework because it does not necessitate the use of any specific resources or libraries. It doesn't have a database abstraction layer, type validation, or any other components that depend on third-party libraries to perform basic tasks. Extensions, on the other hand, may be used to incorporate program functionality as if they were built into Flask itself. Object-relational

mappers, type validation, upload management, various transparent authentication technologies, and other framework-related tools all have extensions. Flask operates based on the following components:

1. **WSGI**: The Web Server Gateway Interface (WSGI) has become the industry standard for developing Python web applications. The Web Service Gateway Interface (WSGI) is a specification for a universal interface between the web server and web applications.

2. **Werkzeug**: It's a WSGI toolkit that handles requests, answers artifacts, and other common tasks. This makes it possible to create a web application on top of it. Werkzeug is one of the foundations of the Flask system.

3. **Jinja2**: Jinja2 is a popular templating engine for Python. A web templating system combines a template with a certain data source to render dynamic web pages.

2.5 Conclusion

In this chapter, researches conducted by the researchers about implementing machine learning in the healthcare system to diagnose the disease are discussed. The experimental methods, data collection method, shortcomings and algorithms used for the result as presented in literature are identified. Also suggested machine learning algorithms are also presented. Later, traditional machine learning algorithms that are used in this regard are shortly explained. In the next chapter, the methodology used in this work is explained briefly.

Chapter 3

Methodology

3.1 Introduction

For the past decade, researchers have conducted many types of research to successfully and efficiently implement artificial intelligence in the healthcare system to increase the quality of the clinical procedure. These researches provided very fruitful outcomes and created an impact in the current modern clinical system. Application of Machine Learning and Artificial Intelligence, in the medical science field for diagnosing a patient with diseases.

In the previous chapter, the outcome of the previous researches is represented. The method used in the researches is observed and also limitations are identified. In this chapter, the methodology implemented in this work to acquire the result is briefly explained.

3.2 Implemented Method Overview

Implemented methodology to diagnose patients with Dengue, Malaria and Typhoid requires feature selection process. After careful selection of features from medical books, data collection is done. Symptoms of the above-mentioned disease directly collected from patients admitted at the hospitals, all over Bangladesh. The data collection process followed with data preprocessing step. In this step string data or raw data is converted to numerical data for computation. Data Analysis is done to gain insight into the data collected and remove unimpactful features. Finalizing the final features, data is fed to numerous machine learning models to observe model accuracy against the dataset. Web Application using frontend and backend technology is developed. An API is used to fetch the patient data

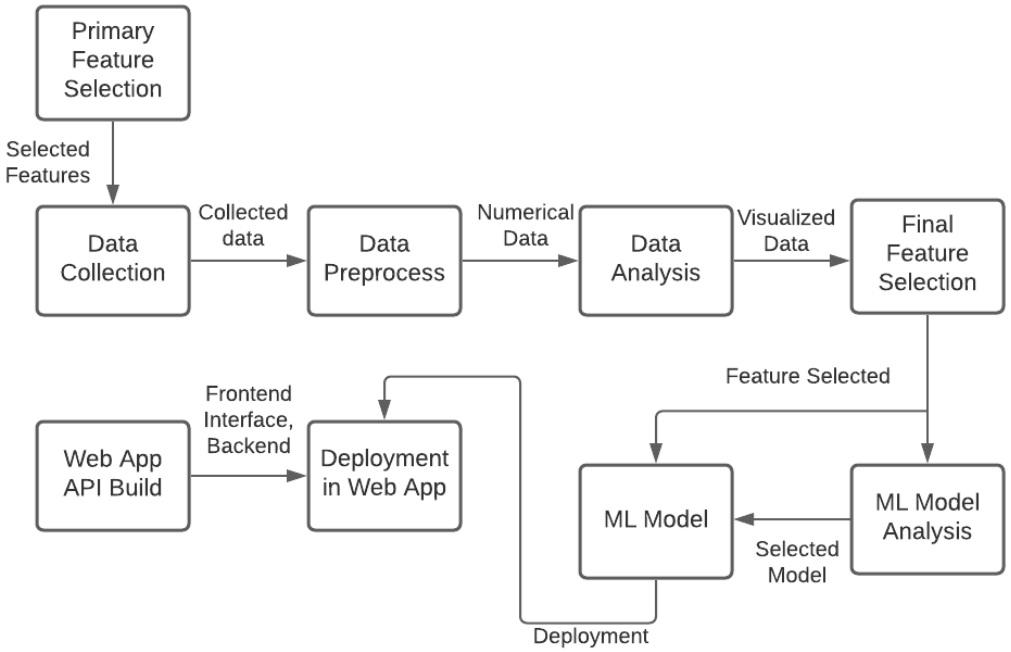


Figure 3.1: Methodology Overview

received from the interface of the app to the backend machine learning model and again render the prediction data from the backend to the frontend interface. The Machine Learning model that provided the best performance is deployed using web technology for high usability.

3.3 Implemented Method

In this section, Implementation method has been briefly described sequentially as presented in the figure 3.1.

3.3.1 Primary Feature Selection

The primary feature selection process is the most crucial part of the whole work. In this case, the primary features are the symptoms of pyrexia patients. Previously researchers have designed and proposed many features that have to be accounted for to diagnose pyrexia patients. But, to differentiate the three most closely symptomized infectious diseases such as Dengue, Malaria, Typhoid, a large

number of features with a cautious selection process is needed as the medical expert suggested [20]. For our work, 29 features are selected to classify the diseases. The features are listed below:

- Age of Patient
- Gender
- Patients Diabetes Condition
- Condition of Blood Pressure
- Fever Type
- Fever Characteristics
- Headache
- Joint Pain
- Muscle Pain
- Fatigue (Tiredness)
- Nausea
- Vomiting
- Sensitivity to Light
- Insomnia
- Abdominal Pain
- Skin Rash
- Stuffy / Runny Nose
- Cough
- Diarrhea
- Sweating
- Constipation
- Dizziness
- Sore Throat
- Back Pain
- Chest Congestion
- Pain in the Eyes
- Poor Appetite
- Convulsions
- Bleeding

3.3.2 Data Collection

The single most critical step in solving any machine learning challenge is data collection. It is, however, a major stumbling block for many academics and computer scientists. Data processing, which mostly consists of data acquisition, data marking, and improving current data or models, typically takes an excessive amount of time.

For this work, the raw data is collected from the patient admitted at the hospitals from five divisions of Bangladesh. The dataset consists of good variance due to

the geographic range of the data collection process. Data collection is done using the questionnaire method. Questions were asked according to the 29 features designed in the first step. No personal information such as Name, Contact Number, Address is registered to keep the data as simple and as anonymous as possible. Each data collection process took 5-10 minutes of work. In this case, patients were queried about the symptoms and effects of the pyrexia they experienced. The information is recorded carefully and labeled instances accordingly.

3.3.3 Data Preprocess

Data Preprocessing is the stage of any Machine Learning process in which the data is translated, or encoded, to make it easier for the machine to process it. In other words, the algorithm can now quickly understand the data's properties. Machines cannot understand raw data. Also, prediction is not achievable by computing string or text data. Again, algorithms need data to be plotted in the graph or plane to classify. Hence, It is necessary to convert the collected raw data to its equivalent numerical data so that machine learning algorithms can be applied to attain the result.

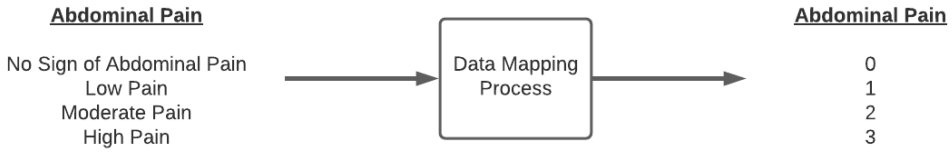


Figure 3.2: Data Mapping

Each raw data collected is mapped to numerical values for computation as shown in 3.2. According to traditional data preprocessing steps, data cleaning is a major part. Many portions of the data may be irrelevant or incomplete. Data cleaning is carried out in order to manage this portion. It entails dealing with incomplete data, noisy data, and so on. But while developing our dataset it is made sure that there is no missing data to ensure its quality. Hence, data cleaning is unnecessary in this regard. As the transformed value ranges are small, normalizing the data won't affect the machine learning performance that much. Therefore, no normalization is done. Dimensionality reduction is also not done.

3.3.4 Data Analysis

Data Analysis is an essential part of data processing. Data Analysis helps the data engineers to select the final features to be used for machine learning algorithms upon which the result is predicted. This step clarifies the relation between features themselves. Also, data analysis helps the data engineer to identify the irrelevant features or variables to drop off. The analysis result assists to understand the story data is trying to tell and determines the relevancy.

The result of the data analyzed using the collected dataset is described in the next chapter, Chapter 4. Data Analysis result is discussed briefly in that section. But the key findings were that all the features selected primarily created an impact on some part of the outcome result. Hence no features are dropped or modified before feeding data to the machine learning model.

3.3.5 Final Feature Selection

When creating a predictive model, feature selection is the method of reducing the number of input variables. The number of input variables can be reduced to minimize the computational expense of modeling and, in some cases, to increase the model's accuracy. The relationship between each input variable and the target variable is evaluated using statistics, and the input variables with the best relationship with the target variable are selected. Since the choice of statistical measurements depends on the data form of both the input and output variables, these approaches can be quick and efficient. The feature selection is done to shortening the training time, avoid the curse of dimensionality and simplification of machine learning model.

To inspect the correlation of primary features we have plotted a correlation matrix to recognize the features that are important also, irrelevant features. Model performance can be increased by dropping off the features that are unimportant in prediction. But in our work, no such features are found hence no feature is dropped or modified.

3.3.6 Model Evaluation and Selection

It's simple to fit a variety of machine learning models on a predictive modeling dataset using efficient machine learning libraries like scikit-learn and Keras. As a result, the complexity of applied machine learning is deciding which model to use for a given problem from a variety of options. Machine learning model selection is a procedure that can be used to compare models of different types as well as models of the same type with different model hyperparameters. The aim of model evaluation is to determine the generalization error of the chosen model. A successful machine learning model, obviously, performs well not just on data seen during training, but also on data not seen during training. As a result, before releasing a model into development, we should be reasonably confident that its output would not deteriorate when faced with new data.

For model selection process, several machine learning models have been trained with the collected preprocessed data for this work. Machine Learning algorithms that are trained and evaluated are listed below:

- Support Vector Machine
- Naïve Bayes
- Decision Tree
- Random Forest
- Logistic Regression
- K-Nearest Neighbors

After evaluating performance of the machine learning algorithms, logistic regression model is selected and deployed using WebApp.

3.3.7 WebApp Development

3.3.7.1 Frontend Development

Front-end web development also referred to as client-side development, is the process of creating and writing HTML, CSS, and JavaScript code for a website

or Web application so that a user can view and communicate with it directly. The difficulty with front-end development is that the methods and strategies used to build the front end of a website change all the time, necessitating the developer's continual awareness of how the area evolves. The goal of website design is to ensure that as visitors visit the site, they see content in an easy-to-read and meaningful format. This is compounded even further by the fact that consumers already use a wide range of devices with different screen sizes and resolutions, prompting the designer to consider these factors when creating the web. The Web App that is used to deploy the machine learning model, used a single-page website. An HTML form is implemented to receive information about the patient for diagnosing results. The interface is beautified with CSS.

3.3.7.2 Backend Development

The server side of development, which is mainly concerned with how the web functions, is referred to as backend development. The primary responsibility will be to make modifications and improvements as well as to manage the site's features. A server, an application, and a database are typically included with this form of web creation. Backend developers write the code that communicates database knowledge to the browser. A backend developer is responsible for something that cannot be seen directly, such as databases and servers. Positions for backend developers are sometimes referred to as programmers or web developers.

Python programming with the help of Flask microservice is used to develop the backend of the WebApp. The selected machine learning model from the model evaluation process is trained and saved as a pickle file. The model is then loaded to the backend codebase. There are two API is developed in the backend:- the main page API and the predict API. The information is received using HTML form at the client-side of the WebApp and then fetched using API to the backend when the predict button is pressed. After button activation the app enters the predict API and received patient information is passed through the trained model to get the prediction result. The result is then rendered to the frontend of the app.

3.3.8 Model Deployment

A machine learning model will only start to add value when the ideas it generates are regularly made available to the consumers for whom it was developed. Deployment is the method of taking a qualified machine learning model and making its predictions available to consumers or other programs. Daily machine learning activities like feature engineering, model discovery, and model verification are not the same as deployment. One of the most complex aspects of extracting benefits from machine learning is model implementation. In our work, the machine learning model is deployed using flask microservice and REST API.

3.4 Conclusion

In this chapter, the method implemented to achieve the target of the work is explained and well discussed. The process is initiated with primary feature selection. After careful selection of the feature, data is collected from the patient admitted at the hospital according to the feature selected. The collected data is preprocessed and mapped to a numerical value for computation purposes. The Numerical data is then analyzed for the final feature selection. In this step, no feature is dropped or modified. The selected features and the data is then fed to several machine learning model to evaluate and select the best performed model. A web App is developed and the model is deployed to the webapp for usability. In the next chapter, implementation results are discussed.

Chapter 4

Results and Discussions

4.1 Introduction

The result is a section that contains a summary of a study's key findings, while the discussion interprets the findings for readers and discusses their importance. The exploration and statistical analysis results of the work are described in the Outcome and Discussion section. It summarizes and makes recommendations based on the information gathered. This section assists other researchers in determining what is possible and what should be predicted when a certain technique is used.

In the previous chapter, the method applied to extract the target result is discussed extensively. The data collection process together with the feature selection process and the collected data analysis are also explained. Also, the evaluation of machine learning algorithms is described.

In this chapter, the outcome of the implemented method is represented. The key findings of the method executed and derived information are also presented. Data are displayed with appropriate figures. Contextual analyses are also given for the derived result.

4.2 Dataset Description

The dataset contains symptoms of 153 patients admitted at the hospitals and diagnosed with either Dengue, Typhoid or Malaria. As per feature design, each instance comprises 29 features that create an impact on the prediction result to efficiently classify three closely symptomized infectious diseases.

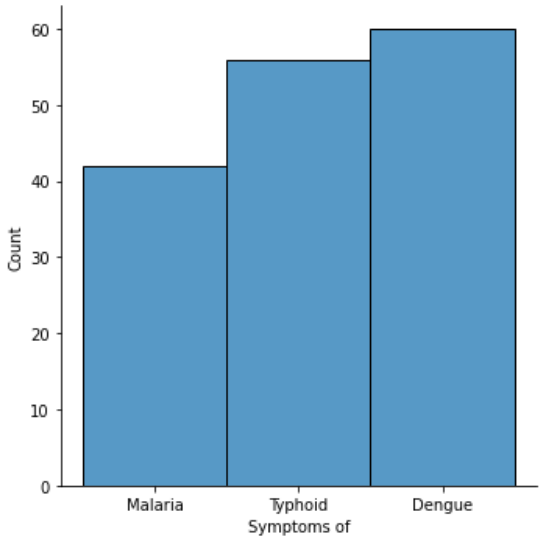


Figure 4.1: No. of Records (Each Class)

Figure 4.1 represents the distribution of 153 data records according to each class. The dataset is minor unbalanced which can be negligible in this case. As it is seen, the dataset contains large records of dengue disease compared to others.

Data was recorded and collected from various divisions of Bangladesh. Since it was collected from a wide geographical region, the dataset contains a good amount of variance in this perspective.

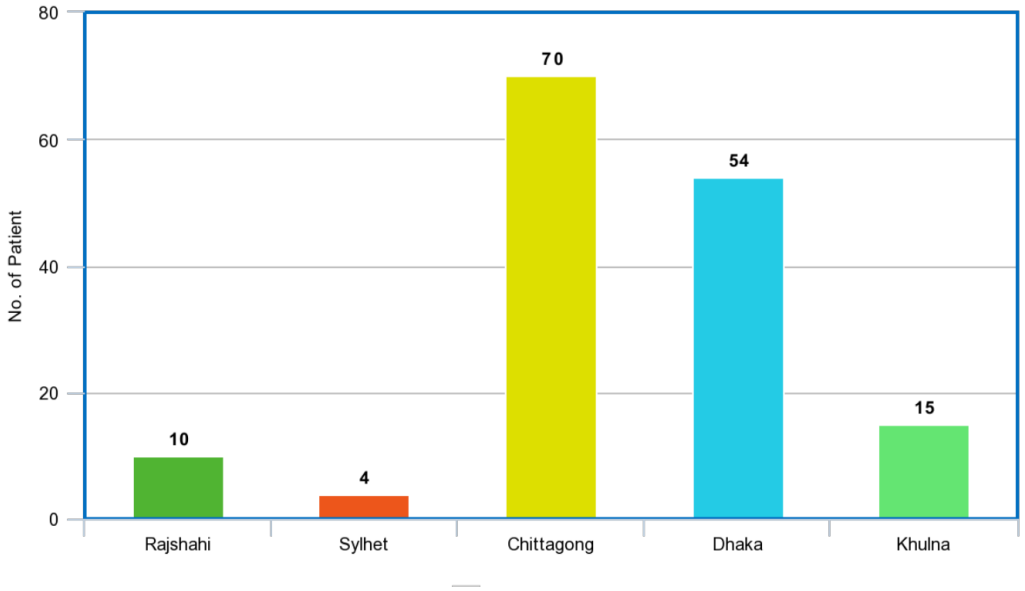


Figure 4.2: No. of Records(Division Wise)

Figure 4.2 shows the number of patient records collected from each division.

Most number of patient records were registered from Chittagong with 70 records, followed by Dhaka division with 54 records.

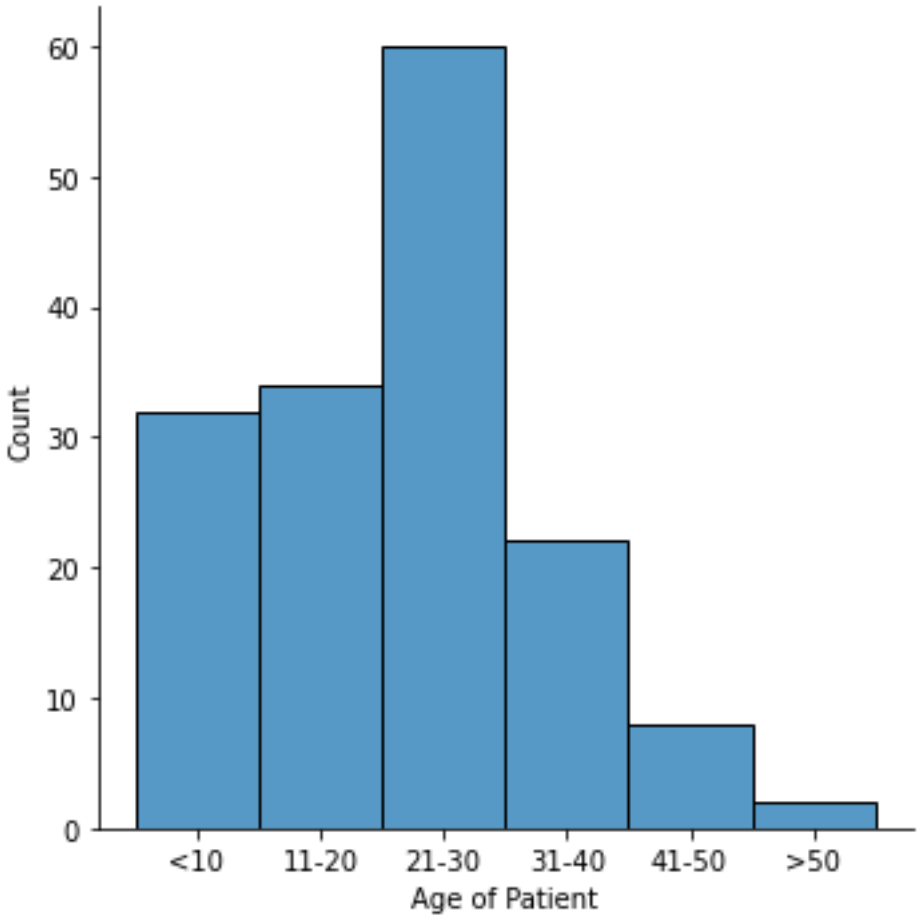


Figure 4.3: No. of Records(Patient Age)

Symptoms were collected from patients of various ages. Figure 4.3 plots the number of the patient against the age range of the patients present in the dataset. The dataset has the highest number of recorded symptoms of the admitted patient aging between 21-30.

For further exploration, in figure 4.4 we have plotted the average age of the patients who have been diagnosed with either dengue, malaria or typhoid and also grouped the data by gender to gain more insight.

In figure 4.5, we have represented the fever characteristics of the patients in terms of percentage. Fever type that a pyrexia patient mostly experience can be observed through this graph chart.

Figure 4.6 explores one of the features used for the machine learning model. In

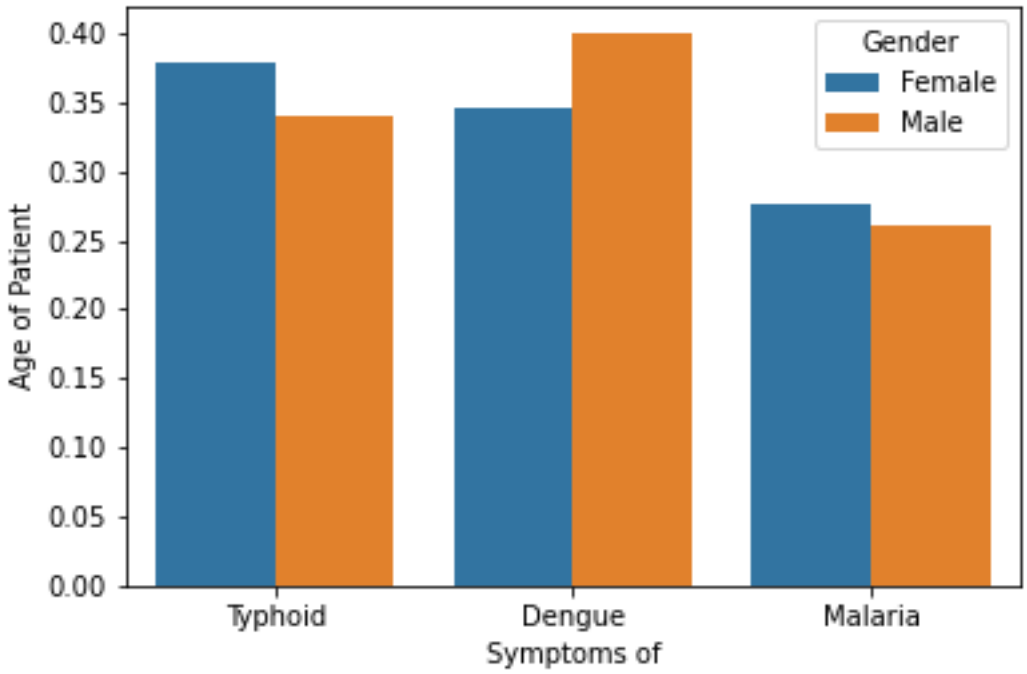


Figure 4.4: Age vs Symptoms

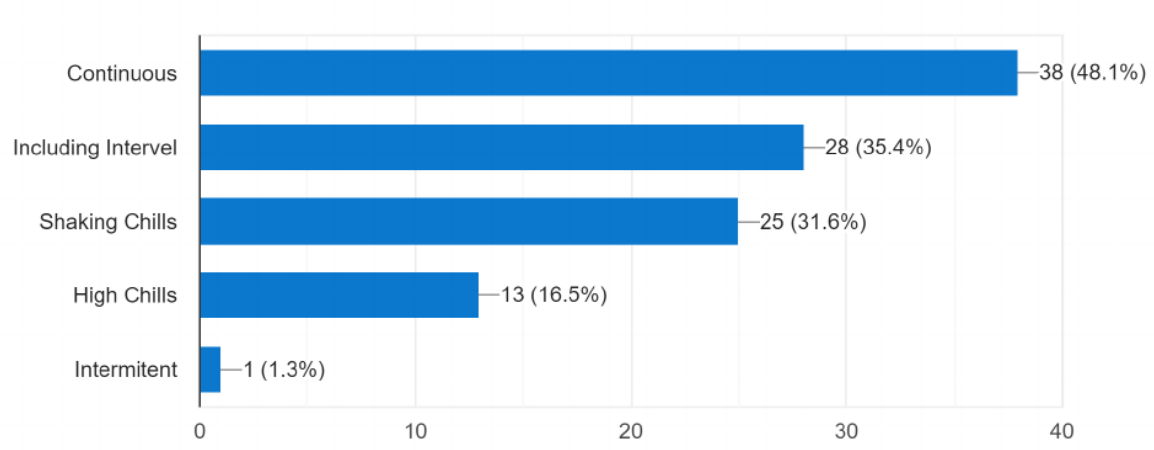
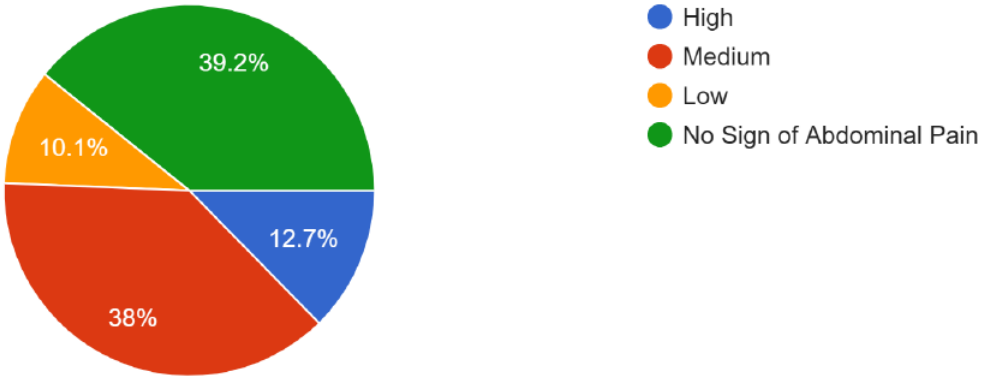


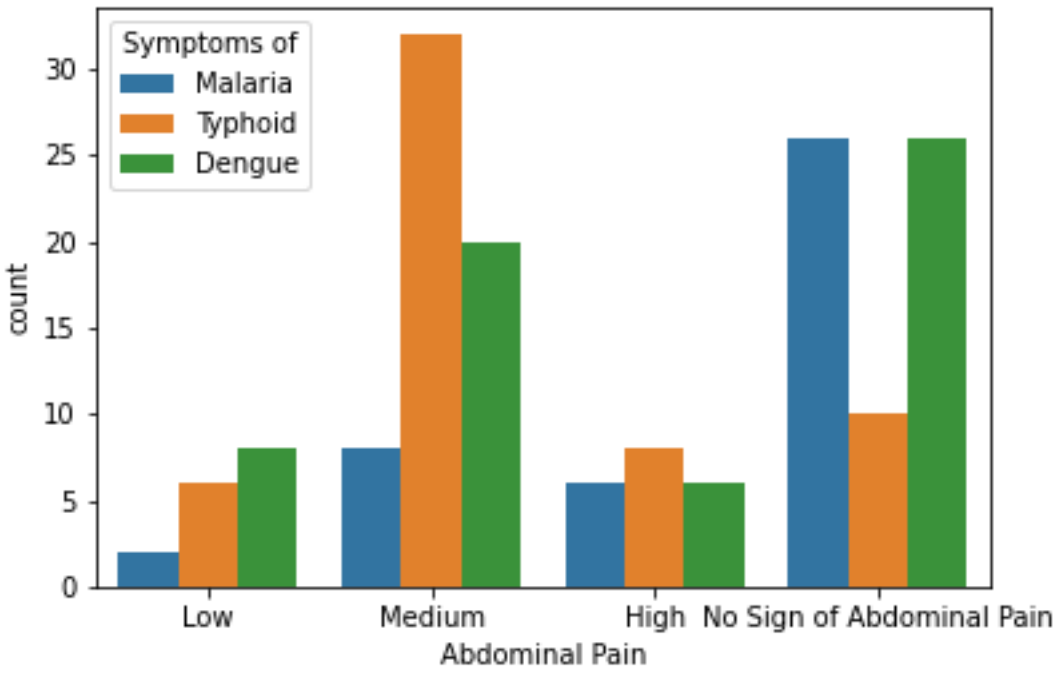
Figure 4.5: Fever Characteristics

subfigure 4.6a abdominal pie chart is plotted where it is seen that among all patients 39.2% of the patients not experienced abdominal pain as the disease symptom. On the other hand, subfigure 4.6b shows the number of patients who experienced or not experienced abdominal pain grouped by the disease diagnosed.

Figure 4.7 explores another one of the core features used for the machine learning model. In subfigure 4.7a blood pressure pie chart is plotted where it is seen



(a) Abdominal Pain Pie Chart

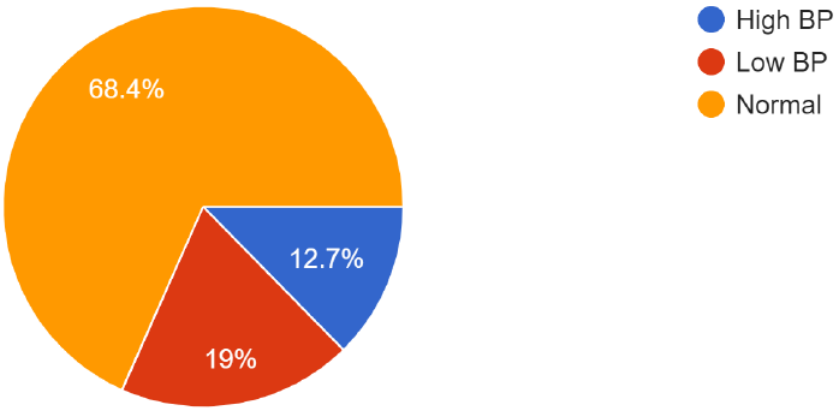


(b) Abdominal Pain grouped by disease

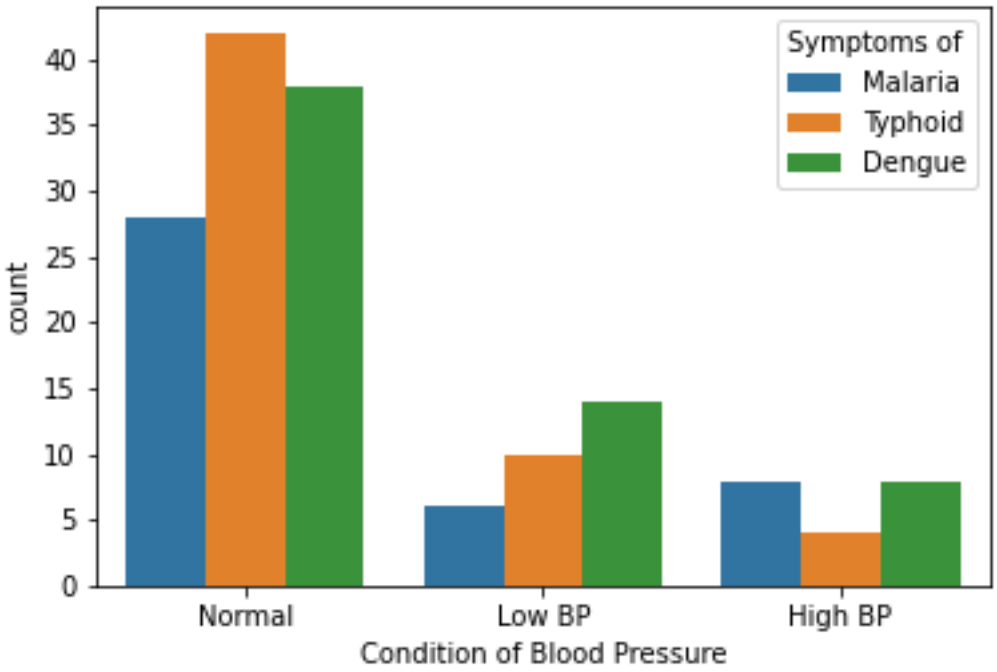
Figure 4.6: Abdominal Pain Exploration

that among all patients 68.4% of the patients not experienced blood pressure as the disease symptom. On the other hand, subfigure 4.7b shows the number of patients who experienced or not experienced blood pressure grouped by the disease diagnosed.

In the figure 4.8 the correlation matrix of the features is represented. As it can be seen that from the correlation matrix, all the features contribute to some extent



(a) Blood Pressure Pie Chart



(b) Blood Pressure grouped by disease

Figure 4.7: Blood Pressure Exploration

to the outcome of the result, no features are dropped or modified for the machine learning model.

Furthermore, we can analyze from the correlation matrix, the topmost key factors that have a higher relation with the target variable. The key factors along with

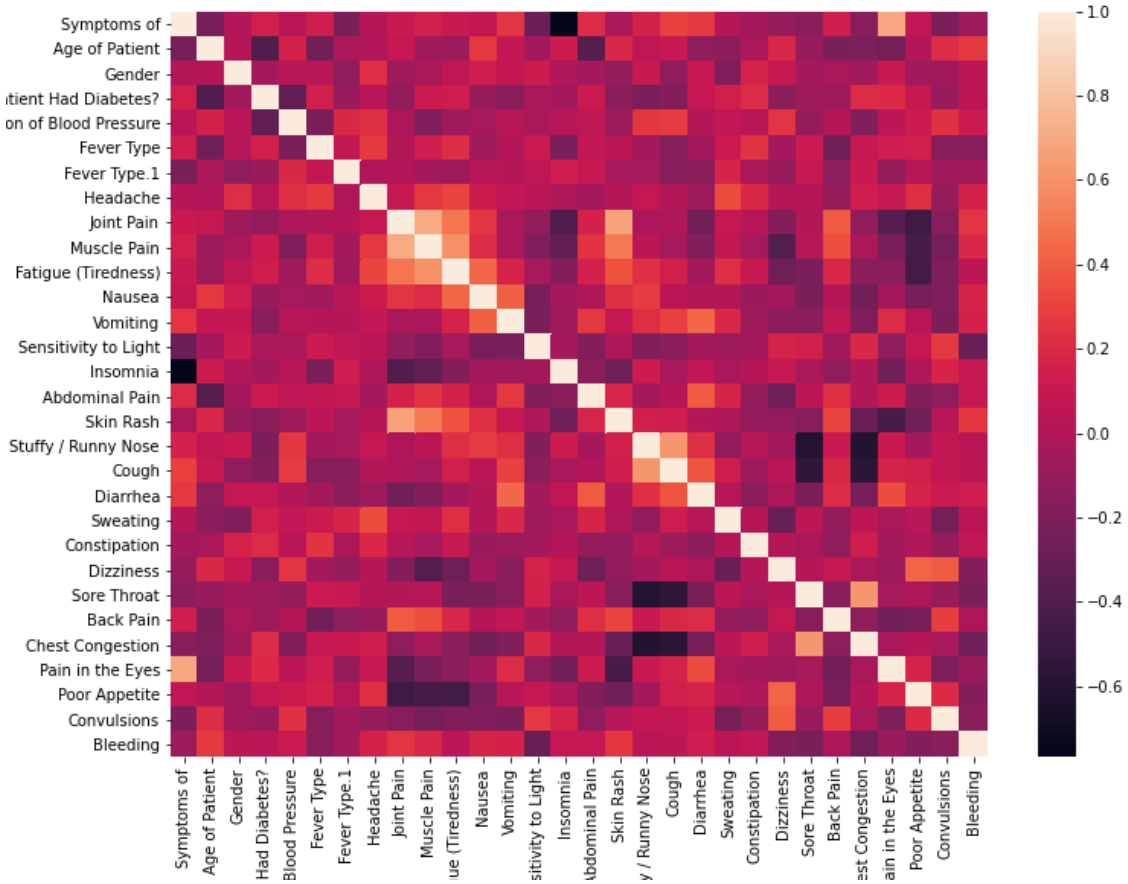


Figure 4.8: Correlation Matrix

Feature	Relation Strength Factor
Insomnia	0.76
Eye Pain	0.71
Cough	0.39

Table 4.1: Key Factors

their relation strength with the output is analyzed and given below:-

4.3 Impact Analysis

Machine Learning is on the brink of becoming a reality. Machine learning is a vast topic of review, and it does not only concern the health industry, but it is affecting every sector one by one in recent years. However, when it comes to the health sector, machine learning is critical for enabling Artificial Intelligence, and the future of healthcare is data-driven. Impact of the work can be quantified as below:

1. A diagnostic support system tool that assists doctors in real-time by retrieving data and making diagnoses recommendations.
2. Machine learning and artificial intelligence can improve the reliability of current disease detection systems.
3. In places where medical services are not accessible, fast diagnostic reports can be collected.

4.4 Evaluation of Performance

The assessment of performance is an important part of the machine learning process. It is, though, a difficult task. As a result, it must be carried out with caution in order to apply machine learning. Any research must have a machine learning algorithm evaluation. As compared to other metrics such as logarithmic loss or any other metric, a machine learning algorithm may produce satisfactory results when measured using an accuracy score metric, but it may produce poor results when compared to other metrics such as logarithmic loss or any other metric. Much of the time, classification accuracy is used to assess a learning model's performance; however, this is insufficient to fully assess a model.

In our work, to compare the performance of machine learning models against our collected and preprocessed dataset we have used the six most popular machine learning classifiers and trained them. For the evaluation process and to determine the best-performed model we have calculated the model accuracy, precision, recall and F1 score of each trained model. For high-quality performance evaluation, we have split the dataset into 75% and 25% for training and testing consecutively. Each of the models was trained with the shuffled 75% of data instances and tested with the other 25%.

Classifier	**Accuracy**	**Precision**	**Recall**	**F1 Score**
SVM	82.5%	0.84	0.82	0.84
Naïve Bayes	93%	0.92	0.93	0.92
K-Neighbors	75.1%	0.74	0.74	0.74
Decision Tree	82.5%	0.90	0.83	086
Random Forest	85%	0.925	0.923	0.923
Logistic Regression	95.1%	0.94	0.94	0.94

Table 4.2: Performance Evaluation of ML Algorithms

As we can analyze from the table 4.2, the logistic regression algorithm provides the best performance in this case with an accuracy of 95.1%. The algorithm also presented promising results with high precision, recall and f1 score.

In previous research work, Sajana et al. [27] implemented a hybrid solution to diagnose only typhoid and dengue where in this work a wide range of features are used and able to diagnose three infectious diseases with high accuracy. Also, in healthcare research, Chen et al[24] designed 26 features to only classify the pyrexia severity level of another infectious disease, COVID-19 and acquired 90% accuracy compared to which our work diagnose three infectious diseases with higher accuracy with our designed 29 features Similar works are also conducted where the severity of typhoid disease [20] and severity of dengue disease [21] is predicted only. Also, where various survey [23] suggested RST, SVM algorithm for this kind of disease diagnosis, our work found logistic regression to be the best model in this case.

4.5 Model Justification

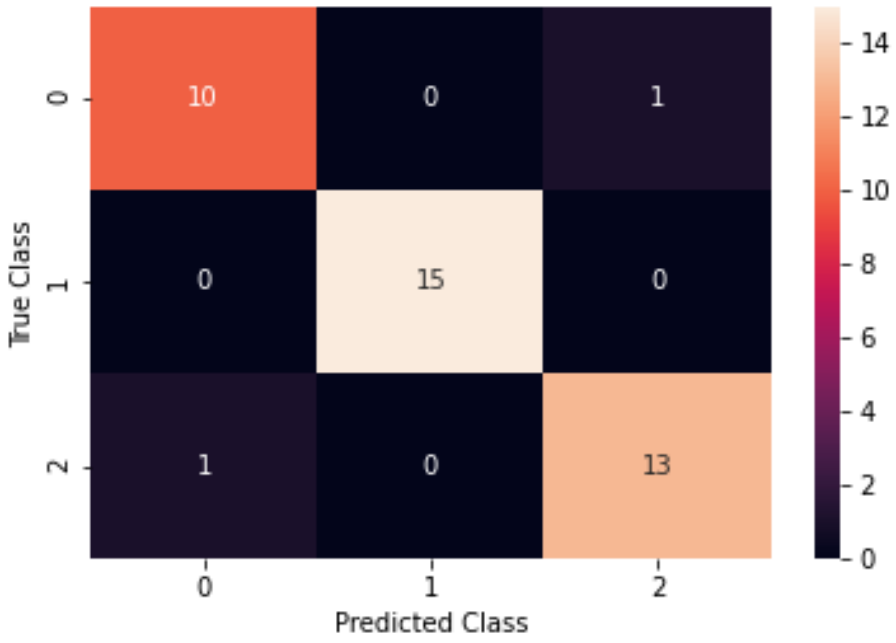
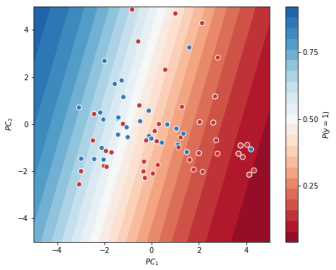


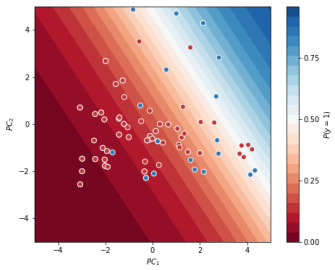
Figure 4.9: Confusion Matrix (Logistic Regression)

In this section we discussed the performance of the logistic regression classifier

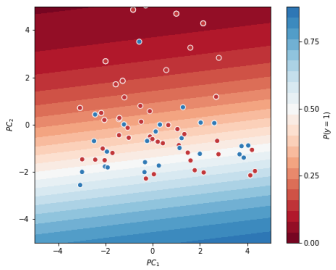
on our test dataset and tried to visualize the reason for logistic regression to be the best algorithm in this regard. In Fig 4.9 confusion matrix of the test cases is plotted for error analysis purpose. Here, class 0, class 1, class 2 represents Dengue, Malaria, Typhoid respectively. As it is seen from the figure, out of 40 test cases only 2 of the cases were misdiagnosed. One Dengue case is predicted as Malaria and for the other Malaria case is predicted as Dengue. Considering the fact that diseases are closely symptomized, the number of false predictions is relatively very low.



(a) Dengue vs Not Dengue



(b) Malaria vs Not Malaria



(c) Typhoid vs Not Typhoid

Figure 4.10: Logistic Regression Model Performance Analysis

Logistic Regression is a statistical method and a linear classifier for binary classification. For multiclass classification, logistic regression adopts the One vs All Method to classify each class. It is impractical to visualize logistic regression performance for 29-dimensional data. However, to understand the classification, the 29-dimension data is reduced to 2 dimensions using the compression technique, Principal Component Analysis. In Fig 4.10(a) we plotted Principle-Component-1 and Principle-Component-2 with logistic regression decision boundary to observe the internal performance of the model to classify Dengue and Not Dengue cases. In this case, Malaria and Typhoid cases are considered as same, the 'Not Dengue' class. As it is seen, Similarly, in Fig 4.10(b) and 4.10(c) we plotted the logistic

regression decision boundary for diagnosing Malaria and Typhoid Cases. Examining Fig 4.10 it can be said that although the figure is useful to envision the machine learning model functioning, but there are several false prediction cases seen in each sub figures. Though blue points should be in the blue region and red points should be in the red region, around 28% of the datapoints are misplaced and residing in the wrong side of the white decision boundary line in all the subfigures. The reason can be discussed using the Table 4.3.

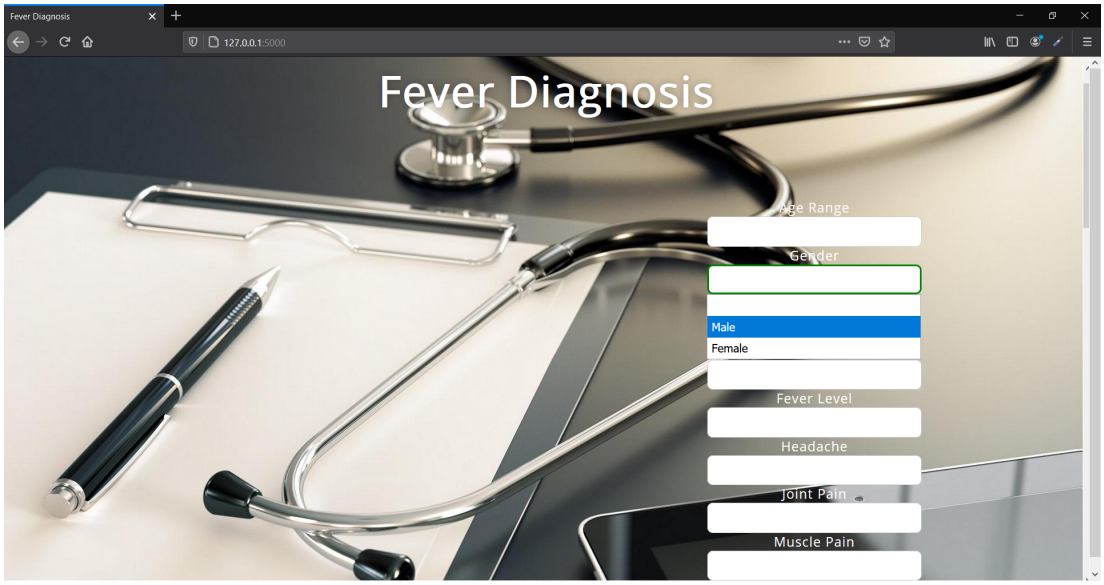
Table 4.3: Variance Ratio of Principal Components

Principal Component	**Explained Variance Ratio(EVR)**	**Cumulative EVR**
PC-1	0.141	0.141
PC-2	0.135	0.276
PC-3	0.095	0.371
PC-4	0.064	0.435
PC-5	0.061	0.496
PC-6	0.059	0.555
PC-7	0.052	0.607
PC-8	0.045	0.652
PC-9	0.043	0.695
PC-10	0.034	0.729
PC-11	0.033	0.762
PC-12	0.027	0.789
PC-13	0.025	0.814
PC-14	0.024	0.838
PC-15	0.021	0.859

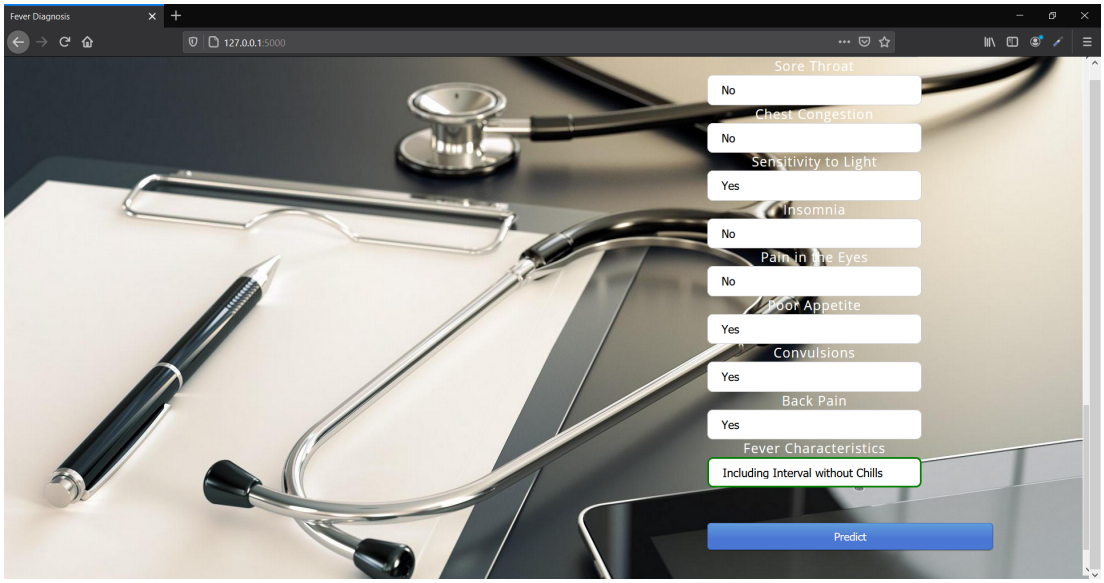
Explained Variance Ratio of the first 15 principle components are calculated using eigenvectors and eigenvalues and listed in Table 4.3. From the table, the first two principal components retains barely 30% of the original information, losing over 70%. With conserving only 30% information logistic regression model classified the data admirably in the Fig 4.10 causing the effective performance of logistic regression in our dataset. To obtain and visualize good result, at least 80% of the information has to be reserved [35]. In that case, at least first 13 principal components has to be plotted which is not possible to visualize in this case. This experimental outcome verifies another analysis result that we observed earlier from correlation matrix that all features are densely correlated, which was the very reason that none of the features were dropped or modified before training the machine learning model.

4.6 Web Application

For the greater usability, the selected model is deployed using WebApp. The WebApp is developed using Flask microservice.



(a) Interface Part-1



(b) Interface Part-2

Figure 4.11: WebApp Interface

Figure 4.11a is the upper half and Figure 4.11b is the lower half of the interface. The Interface is a HTML form that receives the symptoms to diagnose the patients disease. After filling up the form the interface sends the information to the backend and inferred result is shown on the interface.

Figure 4.12: Inferred Result

4.7 Conclusion

The predicted performance of a model can, in theory, tell us how well it performs on unknown data. The biggest thing we try to address is always making forecasts based on future results. Since each machine learning algorithm attempts to solve a problem with a different goal using a different dataset, it's important to consider the context before selecting a metric.

In this chapter, we have explained the result that has been produced by using our methodology. At the start of the chapter explanation of the dataset is given. Also, the Data analysis result and key findings from the data are presented. Next, the model evaluation result and our discovery from the work are represented. Furthermore, a little comparison of our work with the works of the previous researchers is given. The figure of the interface that is designed to increase usability is also presented. In the next chapter, we conclude our work with a summary.

Chapter 5

Conclusion

5.1 Conclusion

In healthcare, machine learning approaches make use of the growing volume of data to optimize patient outcomes. These strategies have a lot of potentials, but they also have a lot of drawbacks. Medical imaging, natural language interpretation of medical records, and common knowledge are the three primary fields where machine learning is used. Many of these fields are concerned with the diagnosis, identification, and forecasting. Today, a vast infrastructure of medical instruments provides data, but there is also no enabling infrastructure in order to efficiently use the data. Health information comes in a variety of formats, which may complicate data formatting and increase noise. According to Andrew Beck of Harvard University, AI will minimize misdiagnosis by up to 85%. According to Stanford University, it will offer general doctors the same degree of precision as expert dermatologists. Researchers at Mayo Clinic have used artificial intelligence to classify the genetic material in brain tumors without the need for a biopsy. Machine learning was able to identify features of MRI scans that doctors were unable to see. Researchers at Stanford are developing an artificial intelligence neural network that can identify skin cancer lesions with the precision of a dermatologist. Using this equipment in any primary care office could offer low-cost early skin cancer screening to the masses. The same deep learning technology is being used in pathology to train an AI program to identify liver lesions. This may be very beneficial to pathologists who deal with patients on a regular basis. Also in genomics, this technique is decoding the 3 billion-long genome sequence of a person right before our eyes. AI would have an effect on everything from prescription development to identifying health insurance fraud,

empowering healthcare providers to concentrate on what matters most: diagnosis and treatment. The effects of using deep learning and AI in healthcare can be felt in a variety of areas that impact patient lives, including medication development, diagnosis of medical photos, treatment plan targeting, and more.

In our work, we tried to create an impact on the modern healthcare system by providing a methodology to diagnose human pyrexia diseases. With the aim of improvement of current clinical system, the work is done.

In the first chapter, the thesis has been given a high-level outline. The reader was introduced to similar topics in this chapter. The difficulty encountered and the implementation of this work are also briefly explored. In the inspiration portion, the importance of the work is discussed, as well as the contribution of the work.

In the next chapter, the researchers' study into applying machine learning in the healthcare sector for disease diagnosis is addressed. The experimental techniques, data collection process, flaws, and algorithms used in the literature for the results are listed. Machine learning algorithms that have been proposed are also discussed. Traditional machine learning algorithms that are used in this context will be briefly discussed later.

In chapter 3, the approach used to accomplish the work's goal is well explained and debated. The method begins with the collection of primary features. Data is obtained from the patient admitted to the hospital-based on the function chosen after due consideration. For computing purposes, the obtained data is preprocessed and mapped to a numerical value. After that, the numerical data is analyzed to determine the final function set. No features are removed or changed in this process. The data and features are then fed into multiple machine learning models, which are then evaluated and the best performing model is chosen. A web app is developed, and the concept is deployed to it for usability testing.

In chapter 4, we've clarified how our methods yielded the results we did. The dataset is explained at the beginning of the chapter. Also, the results of the data collection and main conclusions from the data are discussed. Following that, we present the model validation outcome as well as our findings from the research.

A brief overview of our work with the work of previous researchers is also given. A figure of the interface is also seen and is intended to improve usability.

5.2 Future Work

There are many works that can be done in the future. They are summarized below:

1. **Dataset Quantity:** The dataset contains only 153 instances. The analysis result would have been more accurate if more data is available for the process.
2. **Dataset Variance:** Not only the quantity of the data instances should be increased but data should have collected from a wide range of areas to add diversity.
3. **Applying RL:** Enforcing Reinforcement Learning agent instead of traditional machine learning model can be done.

References

[1] J. L. Jameson, A. S. Fauci, D. L. Kasper, S. L. Hauser, D. L. Longo and J. Loscalzo, *Harrisons principles of internal medicine*. McGraw-Hill Education, 2018 (cit. on p. 1).

[2] Y. K. Axelrod and M. N. Diringer, 'Temperature management in acute neurologic disorders,' *Neurologic clinics*, vol. 26, no. 2, pp. 585–603, 2008 (cit. on p. 1).

[3] K. B. Laupland, 'Fever in the critically ill medical patient,' *Critical care medicine*, vol. 37, no. 7, S273–S278, 2009 (cit. on p. 1).

[4] P. Kiekkas, D. Aretha, N. Bakalis, I. Karpouhtsi, C. Marneras and G. I. Baltopoulos, 'Fever effects and treatment in critical care: Literature review,' *Australian critical care*, vol. 26, no. 3, pp. 130–135, 2013 (cit. on p. 1).

[5] J. E. Sullivan, H. C. Farrar *et al.*, 'Fever and antipyretic use in children,' *Pediatrics*, vol. 127, no. 3, pp. 580–587, 2011 (cit. on p. 1).

[6] *6 most common diseases during monsoon: Dengue, malaria, typhoid, cholera and more- symptoms and prevention*. [Online]. Available: `https://www.timesnownews.com/health/article/6-most-common-diseases-during-monsoon-dengue-malaria-typhoid-cholera-and-more-symptoms-and-prevention/456319` (cit. on p. 1).

[7] T. Vos, C. Allen, M. Arora, R. M. Barber, Z. A. Bhutta, A. Brown, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen *et al.*, 'Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the global burden of disease study 2015,' *The lancet*, vol. 388, no. 10053, pp. 1545–1602, 2016 (cit. on p. 1).

[8] H. Wang, M. Naghavi, C. Allen, R. M. Barber, Z. A. Bhutta, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen, M. M. Coates *et al.*, 'Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the global burden of disease study 2015,' *The lancet*, vol. 388, no. 10053, pp. 1459–1544, 2016 (cit. on p. 1).

[9] W. H. Organization *et al.*, 'Typhoid vaccines: Who position paper, march 2018–recommendations,' *Vaccine*, vol. 37, no. 2, pp. 214–216, 2019 (cit. on p. 1).

[10] ——, 'Dengue and severe dengue,' World Health Organization. Regional Office for the Eastern Mediterranean, Tech. Rep., 2014 (cit. on p. 1).

[11] G. A. Roth, D. Abate, K. H. Abate, S. M. Abay, C. Abbafati, N. Abbasi, H. Abbastabar, F. Abd-Allah, J. Abdela, A. Abdelalim *et al.*, 'Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: A systematic analysis for the global burden of disease study 2017,' *The Lancet*, vol. 392, no. 10159, pp. 1736–1788, 2018 (cit. on p. 1).

[12] W. H. ORGANIZATION, *WORLD MALARIA REPORT 2019*. WORLD HEALTH ORGANIZATION, 2019 (cit. on p. 1).

[13] J. J. Waggoner, L. Gresh, M. J. Vargas, G. Ballesteros, Y. Tellez, K. J. Soda, M. K. Sahoo, A. Nuñez, A. Balmaseda, E. Harris *et al.*, 'Viremia and clinical presentation in nicaraguan patients infected with zika virus, chikungunya virus, and dengue virus,' *Clinical Infectious Diseases*, ciw589, 2016 (cit. on p. 1).

[14] R. Bhardwaj, A. R. Nambiar and D. Dutta, 'A study of machine learning in healthcare,' in *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, IEEE, vol. 2, 2017, pp. 236–241 (cit. on p. 1).

[15] J. Wiens and E. S. Shenoy, 'Machine learning for healthcare: On the verge of a major shift in healthcare epidemiology,' *Clinical Infectious Diseases*, vol. 66, no. 1, pp. 149–153, 2018 (cit. on p. 1).

[16] A. Callahan and N. H. Shah, 'Machine learning in healthcare,' in *Key Advances in Clinical Informatics*, Elsevier, 2017, pp. 279–291 (cit. on p. 1).

[17] S. Ralston, M. W. J. Strachan, R. Britton, I. D. Penman and R. P. Hobson, *Davidsons principles & practice of medicine*. Elsevier, 2018 (cit. on p. 3).

[18] T. M. Mitchell, *Machine learning*. MacGraw-Hill, 1997 (cit. on p. 3).

[19] M. Atul Kaushal, M. Ken Abrams, M. David Sklar and M. Bill Fera, *The future of artificial intelligence in health care*, Dec. 2019. [Online]. Available: `https://www.modernhealthcare.com/technology/future-artificial-intelligence-health-care` (cit. on p. 4).

[20] A. Oguntimilehin, A. Adetunmbi and O. Abiola, 'A machine learning approach to clinical diagnosis of typhoid fever,' *A Machine Learning Approach to Clinical Diagnosis of Typhoid Fever*, vol. 2, no. 4, pp. 1–6, 2013 (cit. on pp. 6, 15, 29).

[21] C. Davi, A. Pastor, T. Oliveira, F. B. de Lima Neto, U. Braga-Neto, A. W. Bigham, M. Bamshad, E. T. Marques and B. Acioli-Santos, 'Severe dengue prognosis using human genome data and machine learning,' *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2861–2868, 2019 (cit. on pp. 7, 29).

[22] S. Srivastava, S. Soman, A. Rai and A. S. Cheema, 'An online learning approach for dengue fever classification,' in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, IEEE, 2020, pp. 163–168 (cit. on p. 7).

[23] M. Fatima, M. Pasha *et al.*, 'Survey of machine learning algorithms for disease diagnostic,' *Journal of Intelligent Learning Systems and Applications*, vol. 9, no. 01, p. 1, 2017 (cit. on pp. 7, 29).

[24] Y. Chen, L. Ouyang, F. S. Bao, Q. Li, L. Han, B. Zhu, Y. Ge, P. Robinson, M. Xu, J. Liu *et al.*, 'An interpretable machine learning framework for accurate severe vs non-severe covid-19 clinical type classification,' *Available at SSRN 3638427*, 2020 (cit. on pp. 7, 29).

[25] Y. W. Lee, J. W. Choi and E.-H. Shin, 'Machine learning model for predicting malaria using clinical information,' *Computers in Biology and Medicine*, vol. 129, p. 104 151, 2021 (cit. on p. 8).

[26] A. Oguntimilehin, A. Adetunmbi and K. Olatunji, 'A machine learning based clinical decision support system for diagnosis and treatment of typhoid fever,' *A Machine Learning Based Clinical Decision Support System for Diagnosis and Treatment of Typhoid Fever*, vol. 4, no. 6, pp. 1–9, 2014 (cit. on p. 8).

[27] T. Sajana, M. Syamala, L. P. Maguluri and C. U. Kumari, 'A hybrid approach for classification of infectious diseases,' *Materials Today: Proceedings*, 2021 (cit. on pp. 8, 29).

[28] C. Cortes and V. Vapnik, 'Support-vector networks,' *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995 (cit. on p. 8).

[29] A. Ben-Hur, D. Horn, H. T. Siegelmann and V. Vapnik, 'Support vector clustering,' *Journal of machine learning research*, vol. 2, no. Dec, pp. 125–137, 2001 (cit. on p. 8).

[30] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, 'Top 10 algorithms in data mining,' *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008 (cit. on p. 9).

[31] A. McCallum, 'Graphical models, lecture2: Bayesian network represention,' *PDF). Retrieved*, vol. 22, 2019 (cit. on p. 9).

[32] E. Fix, *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF school of Aviation Medicine, 1985, vol. 1 (cit. on p. 9).

[33] J. Tolles and W. J. Meurer, 'Logistic regression: Relating patient characteristics to outcomes,' *Jama*, vol. 316, no. 5, pp. 533–534, 2016 (cit. on p. 10).

[34] T. K. Ho, 'Random decision forests,' in *Proceedings of 3rd international conference on document analysis and recognition*, IEEE, vol. 1, 1995, pp. 278–282 (cit. on p. 10).

[35] I. Lindgren, *Dealing with highly dimensional data using principal component analysis (pca)*, Apr. 2020. [Online]. Available: `https://towardsdatascience.com/dealing-with-highly-dimensional-data-using-prin`

cipal-component-analysis-pca-fea1ca817fe6%5C#:~:text=The%20explained%20variance%20ratio%20is,or%2080%5C%%20to%20avoid%20overfitting. (cit. on p. 31).