PROJECT PROGRESS REPORT

on

SEPSIS DETECTION

Submitted by
Yash Puri - 2000290120196
Yash Goel - 2000290120195
Vishal Verma - 2000290120193

Under the guidance of
Ms. Neha Shukla (Assistant Professor CS)

Submitted to the department of Computer Science in partial

fulfilment of the requirements

for the degree of

Bachelor of Technology

in

Computer Science

KIET GROUP OF INSTITUTIONS, Ghaziabad

Dr. A.P.J. Abdul Kalam Technical University, Uttar Pradesh Lucknow

DECLARATION

We hereby declare that this submission is our own work that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature:
Name: Yash Puri
Roll number: 2000290120196
Date:

Signature:
Name: Yash Goel
Roll number: 2000290120195
Date:

Signature:
Name: Vishal Verma
Roll number: 2000290120193
Date:

CERTIFICATE

This is to certify that Project Report entitled "**SEPSIS DETECTION**" which is submitted by **Yash Puri, Yash Goel and Vishal Verma** in partial fulfillment of the requirement for the award of degree B. Tech. in Department of Computer Science & Engineering of Dr. A.P.J. Abdul Kalam Technical University, formerly Uttar Pradesh Technical University, is a record of the candidate's own work carried out by them under my supervision.

(Supervisor Signature)

Name: Ms. Neha Shukla

Designation: Assistant Professor CS

Date:

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Ms. Neha Shukla, Assistant Professor, Department of Computer Science for her constant support and guidance throughout the course of our work. His sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of Professor Dr. Ajay Kumar Shrivastava, Head, Department of Computer Science, KIET Ghaziabad for his full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature: Signature:

Name: Yash Puri   Name: Yash Goel

Roll No: 2000290120196   Roll No: 2000290120195

Date: Date:

Signature:

Name: Vishal Verma

Roll No:

2000290120193

Date:

ABSTRACT

This abstract provides an overview of sepsis, symptoms and diagnostics and also about the objectives of developing sepsis prediction model which will act as an aid for healthcare professionals for its early detection. Sepsis is activated by the immune system present in our body that works all the time in order to prevent the infection from entering. During this stage, the enormous number of synthetic substances discharged into the blood causes broad irritation. Sepsis occurs when body's response to the chemicals is out of balance, triggering changes that can damage multiple organ substances. Timely detection and interventions are very crucial for improving patient outcomes as sepsis carries a high mortality rate if left untreated. For the patient the practicality of predicting sepsis disease occurrence in development is an important factor in the result. The primary goal is to build models using different machine learning algorithms such as Logistic Regression, KNN, Naive Bayes and Random Forest and then to find out the best classifier with high accuracy and detect the sepsis disease in minimal time where Random Forest performed best among all other considered for predicting sepsis at an early stage with an accuracy of almost 96 percent. Our secondary goal is to build and design a user-friendly web application.

LIST OF FIGURES

LIST OF ABBREVIATIONS

S.No	Abbreviations	Full Form
1.	ML	Machine Learning
2.	OS	Operating System
3.	KNN	K-Nearest Neighbor
4.	WI	Web Interface
5.	AI	Artificial Intelligence
6.	RF	Random Forest
7.	URL	Uniform Resource Locator

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION   Page No.

1.1 Introduction to Project

1.2 Project Category

CHAPTER 2: LITERATURE REVIEW

2.1 Literature Review

2.2 Problem Formulation

2.3 Objectives

CHAPTER 3: PROPOSED SYSTEM

3.1 Proposed System

3.2 Unique Features of The System (Difference from Existing System)

3.3 Structure of Report

CHAPTER 4: (REQUIREMENT ANALYSIS AND SYSTEM SPECIFICATION)

4.1 Feasibility Study (Technical, Economical, Operational)

4.2 Software Requirement Specification Document Which Must Include the Following:

(Data Requirement, Functional Requirement, Performance Requirement, Maintainability Requirement, Security Requirement)

4.3 SDLC Model to Be Used

4.3.1 System Design Using DFD Level 0 And Level 1

4.3.2 Use Case Diagram

4.3.3 Database Design

4.3.4 ER Diagram

CHAPTER 5:   IMPLEMENTATION

5.1 Introduction to Languages, Tools and Technologies Used for Implementation.

CHAPTER 6: TESTING, AND MAINTENANCE

6.1 Testing Techniques and Test Cases Used

CHAPTER 7: (RESULTS AND DISCUSSIONS)

7.1 User Interface Representation (Of Respective Project)

7.2 Brief Description of Various Modules of The System

7.3 Snapshots of System with Brief Detail of Each

7.4 Back Ends Representation (Database to Be Used)

7.5 Snapshots of Database Tables with Brief Description

CHAPTER 8: CONCLUSION AND FUTURE SCOPE

CHAPTER 1

INTRODUCTION

1.1 INTRODUCTION TO PROJECT

Interesting but as well as difficult for predicting clinical problems to improve the patient outcomes hence predicting clinical diseases effectively has always been a critical area to concentrate on.

"The sepsis prediction can be viewed as an accurate guess regarding the patients onset of sepsis presence or absence over a specific time period without causing any delays, based on the set of vital parameters provided by the doctor to predict sepsis at an early stage."

Sepsis is a hazardous condition that happens when the body's reaction to contamination causes tissue harm, organ failure, or even demise of the person. The body releases natural synthetics into the blood circulation system in order to counterbalance the infection which is present inside. Sepsis occurs when the body's response to these chemicals is out of balance, this can damage many organ systems. It is most common dangerous for senior citizens, pregnant ladies, kids below one-year-old, persons suffering from chronic conditions, such as diabetes, kidney disease, lung disease, or even cancer, as they have weak immune systems. Several examinations have demonstrated that delays in finding and treatment of sepsis can prompt high death rates.

Earlier sepsis prediction was used to be a time consuming task and may cause delays in treatment which was done through traditional methods. But in today's fast pacing and competitive world, where competitive margins are shrinking and decisions needed to be quick but well-informed, an effective and efficient predicting system has become necessary not only for sepsis or clinical diseases to improve the patient outcomes but rather for all business. The, manual execution of this task could result in significant errors, resulting in poor results of treatment, and, more precisely, will take lot of time, which is not preferable in today's fast-paced environment. Machine learning is the field in which computers learn to surpass humans in specific tasks. They are utilized to perform specific jobs in a rational manner in order to achieve better results for the advancement of present society.

In this paper, the performance of a variety of machine learning models for the application of sepsis prediction at an early stage to act as an aid for health care professionals is measured. And best performing machine learning model is chosen for predicting the onset of sepsis accurately which will ultimately help to improve the patient outcomes.

1.2 PROJECT CATEGORY

The sepsis prediction project must fall under the category of "Application" because it involves development of a predictive model for early prediction of sepsis using advanced technologies such as machine learning for effective and accurate predictions.

CHAPTER 2

LITERATURE REVIEW

2.1 LITERATURE REVIEW

In this project, a literature review has been conducted for the finding suitable prediction model for detecting sepsis at an early stage by looking at past research done in sepsis prediction in health care sectors using different machine learning algorithms.

Reference	Major Contributions	Objective	Year	Domain/ Stack	Result
[1]	Predicting Infections Using Computational Intelligence.	To develop computation models using ml algorithm for predicting sepsis by infections and surgical infections.	2020	Machine Learning	The system learns for predicting sepsis mainly by infections and surgical infections.
[2]	A Deep Learning-Based Sepsis Estimation Scheme.	The objective was to design a machine learning based technique that can predict cases of septic shock and extreme sepsis.	2020	Machine Learning	Constructed a model to predict septic and extreme sepsis.
[3]	Trans-thoracic echocardiography and mortality in sepsis.	To examine the role of TTE with 28- day mortality in population. The MIMIC- III database to identify patients with sepsis who had and had not received TTE.	2018	Machine Learning	Suggested useful diagnostic tool for clinical decision and support.

`

[4]	Learning Representations for the early detection of sepsis with deep neural networks.	To provide an effective early-stage sepsis detection model using deep learning algorithms.	2017	Machine Learning	Provided an effective system for prediction but complex and may cause different results.
[5]	New Effective Machine Learning Framework for Sepsis Diagnosis.	Provide method which has got 81.6% recognition rate, 89.57% sensitivity and used effectively to diagnose sepsis.	2018	Machine learning	It sometimes provides false results resulting in unnecessary treatment.
[6]	Early prediction of sepsis based on ML techniques.	Used XG boost and light gbm to predict sepsis 6 hours in advance.	2021	Machine learning	If the data used is biased it may result in unreliable prediction.
[7]	Multi branching CNN for predicting sepsis.	Multi branching framework to model complex clinical data to predict sepsis.	2021	Machine learning	It becomes complex when the data becomes very large.
[8]	Predicting of sepsis patients using machine learning approach.	Study found that machine learning models performed better than other existing systems.	2019	Machine learning	It doesn't specify that which model is best for predicting sepsis patients.
[9]	Ensemble ML model for early prediction of sepsis.	Proposed an ensemble model by using bagging and boosting trees for predicting sepsis.	2019	Machine learning	It may produce inaccurate predictions.
[10]	Early prediction of sepsis for ICU patients.	Proposed system intended to find and test ml algorithms for prediction.	2022	Machine learning	Implementing to real world may face challenges in terms of acceptance and integration.

TABLE 1

`

The comparison of several research papers which are discussed in table 1 are given below:-

In [1], Baldominos et *al.* stated that machine learning techniques such as SVM, Logistic Regression, Naïve Bayes were effective in detecting infections mainly focused towards predicting sepsis. This study mainly focused on how effectively infections can be predicted including sepsis also by using various machine learning algorithms but its implementation cost was expensive because it may involve restructuring of health care systems. In other study use of some specific machine learning model was emphasized to predict cases of sepsis and septic shock by highlighting the relevance and importance of laboratory values of the affected patients but as layers become more the accuracy becomes less which was proposed by Al-Mualemi in [2]. But later on in a similar study an advance and evolved version of C.N.N was proposed that was multi branching C.N.N a novel predictive framework to predict sepsis

[7] and also handle the missing values and other data issues but it also had a limitation that the framework becomes complex to operate effectively when the data becomes large. A study focused on developing a deep learning based model for predicting sepsis and also achieved better performance with input to output layer by feeding it in one direction [4], but due to the complexity of the model it was difficult for the model to generate same results for sepsis on different datasets which makes it challenging to compareits performance with the others.

In 2019, a study proposed an idea for use of ensemble machine learning model by using boosting and bagging techniques for predicting sepsis [9]. Both techniques are used for correcting errors , optimizing the models performance , enhancing accuracy. Boosting helps to remove and correct errors sequentially that were made by the previous ones whereas bagging technique works for removing the over fitting problem by averaging the results of different machine learning models but this model was highly dependent on the quality of the input data due to which poor quality of input can ultimately result in inaccurate predictions and poor performance of the model. Similarly a study proposed the use of XG boost and light gbm algorithm to predict sepsis 6 hours in advance [6]. Light gbm is also a gradient boosting framework which uses decision trees to increase the performance also it has fast iteration in training which gives the model a better predictive ability to detect sepsis but if the data provided is biased or we can say the data is incomplete then the accuracy of the model can be affected ultimately leading to unreliable results.

In [3], Feng et *al*. it focused on matching the relationship between trans-thoracic echocardiography and patient within septic addressing the use of TTE but performing TTE can be time consuming which can cause delay in the required necessary treatment or prevention.

In [10], The conducted study was intended to propose a system for predicting sepsis by finding and testing out various potential machine learning algorithms such as decision trees, gradient boosted trees but these models were less accurate and may face some challenges in terms of acceptance and integration with healthcare organizations. In a study feature selection was performed on random forest model before building their classification model to optimize it which was then used to effectively diagnose the sepsis but the proposed model has lower specificity rate that means it is more likely to produce false results which will ultimately lead to unnecessary treatment [5]. In a study it was stated that machine learning models are proved to be superior in identifying and predicting sepsis or other infections also over other existing systems and will have very less false results if implemented correctly which can lead to improve the patient outcomes who are suffering from sepsis but the study did not suggested that which machine learning model would be best for predicting sepsis [8].

`

In [11], M. Deepika et al. first of all set up the base for healthcare diagnostics and various limitations present in it also how they can be coped by using machine learning techniques. The author conveys that the systems based on machine learning can be used to find various complications before they occur at very early stage of diagnosis and can be prove to be very useful for treatment and patient outcomes.

As per the papers studied as a part of literature review previous studies that are conducted on the Machine Learning Systems have shown the effectiveness of different machine learning algorithms in different sectors including healthcare.

2.2 PROBLEM FORMULATION

The problem formulation for sepsis detection system project revolves around the need of developing a reliable and accurate system that can identify sepsis at an early stage working as an aid for healthcare providers to focus on taking required actions immediately without having any delays and improve patient outcomes. This system will leverage the use of advance technologies such as machine learning and its algorithms for improved accuracy and efficiency of detecting sepsis. The project aims to enhance the patient outcomes suffering from sepsis, reduce in mortality rates etc. This early detection of sepsis will lead to timely initiation of required treatment and preventions along with reducing the mortality and morbidity rates associated with sepsis.

2.3 OBJECTIVES

All nations, whether underdeveloped, developing, or even developed, rely on healthcare sectors as their primary means of supplying their populations with good treatment. It is very challenging to ensure quality treatment at a large scale without any delays in chronic diseases such as sepsis to improve the patient outcomes. The power of AI and ML can be applied in the field of healthcare as well which can help the doctors to become productive and informed timely for chronic health conditions without any delays.

This forms the main objective and motivation of our project SEPSIS DETECTION SYSTEM.

Through this project we plan to achieve the following objectives –

● To identify the best classifier among Logistic Regression, KNN, Naive Bayes and Random Forest for prediction of sepsis at early stages.
● To design and develop a website and integrate with the model.
● Helping doctors to make smart choices about treatment and prevention to reduce the chances of severe after affects.
● Takes patient data with vital parameters and predict sepsis with accuracy.

`

CHAPTER 3

PROPOSED SYSTEM

3.1 PROPOSED SYSTEM

This chapter includes proposed system for detecting sepsis at an early stage along with its architecture along with modules and data set.
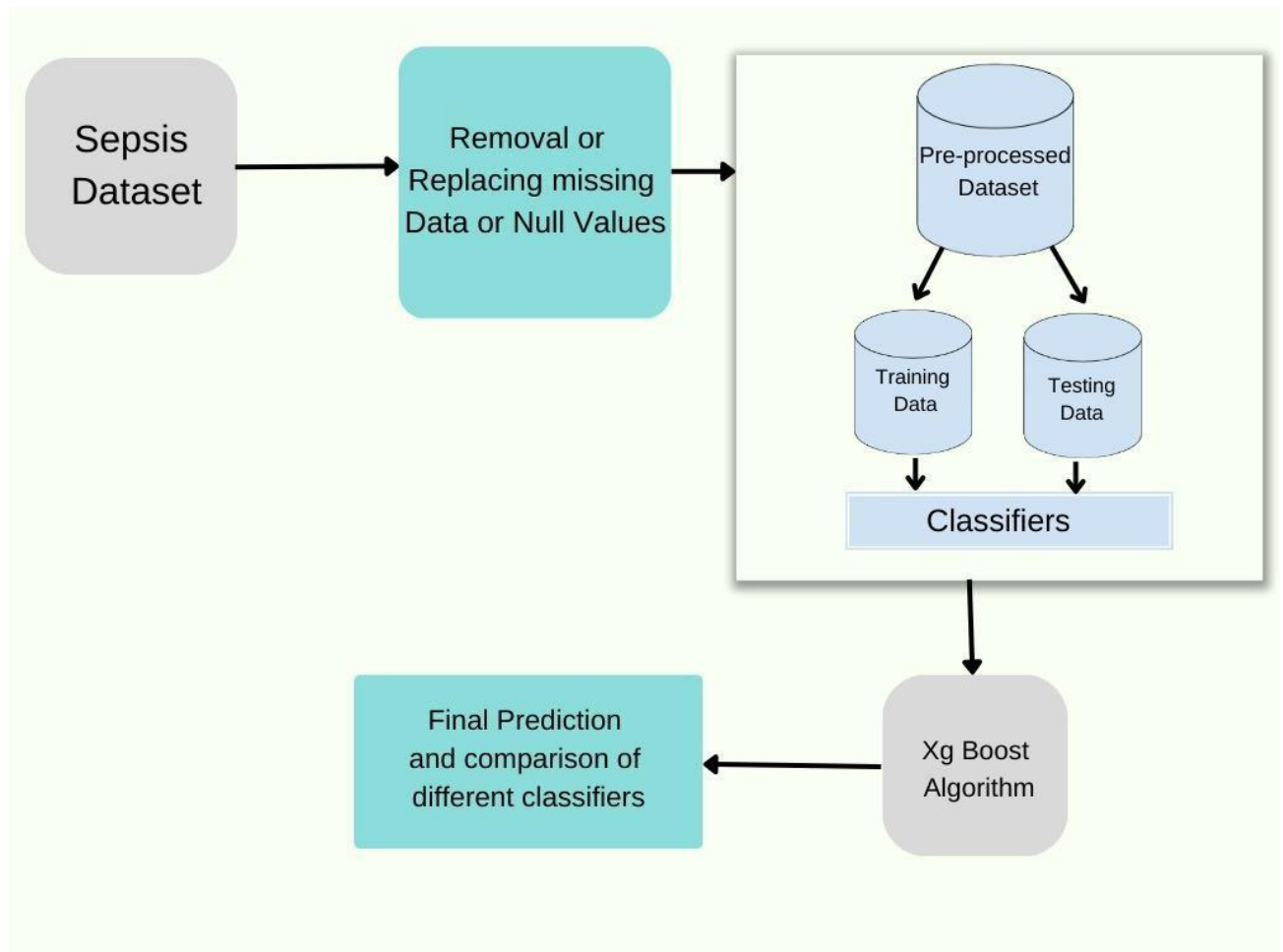
3.1.1 Architecture

3.2 DATASET

In this project Physionet challenge data set is used to identify the best classifier for predicting sepsis at an early stage. The data is gathered from patients of ICU from 3 different hospitals. A total of approximately 40,000 patients clinical data from two definite hospitals were shared with the members. Each patient's clinical data contained likely 40 measurements of vital signs, laboratory, demographics data. Each file has data separated with pipes in which each row represents a 1 hour worth of data. Extremely Imbalance data: The records are extremely imbalanced (More than 97.8% are having 0 sepsis label and 2.2% have sepsis) with the minority class being Sepsis Missing Data:

`

In the data set the percentage of data which is missing is high. This is handled by ignoring the features with more than 80% of missing data.

FEATURES

- Respiratory rate, Temperature, Mean arterial Pressure etc. are Vital Signs.
- Platelet Count, Glucose, Calcium etc. are Laboratory Values.
- Age, Gender, Time in ICU, Hospital Admit time etc. are considered as Demographics.
- 0 (Non-sepsis) and 1 (Sepsis) is the label for identification.

PROPOSED SYSTEM

- Collect our data set.
- Refine it and handle errors.
- Data pre processing.
- Find correlation in data set.
- Define the model.
- Train the model using different algorithms.
- Check model on testing data.
- Model evaluation.
- Choose the best performing model.
- If desirable use this model to predict sepsis in real scenario.

ALGORITHMS USED

Naive Bayes
It is based on the Bayes' theorem notion, which includes estimating the likelihood of an event based on information or knowledge already known. A probabilistic machine learning model called the Naive Bayes classifier is utilized for classification tasks. Naive Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. Naive Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

Random Forest
Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. Random Forest is a popular and effective machine learning technique. Multiple decision trees are combined in this ensemble method to produce predictions.

KNN
One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbour. The K-NN algorithm makes the assumption that the new case and the existing cases are comparable, and it places the new instance in the category that is most like the existing categories. A new data point is classified using the K-NN algorithm based on similarity after all the existing data has been stored. This means that utilising the K-NN method, fresh data can be quickly

`

and accurately sorted into a suitable category. Although the K-NN approach is most frequently employed for classification problems, it can also be utilised for regression.

Logistics Regression

Its ease of use and interpretability make it a common technique in statistics and machine learning. The dependent variable in logistic regression is binary, i.e., it might have two different potential results. For situations involving binary and linear classification, logistic regression is a straightforward and more effective approach. It's a classification model that's incredibly simple to implement and performs admirably with linearly separable classes.

3.2 UNIQUE FEATURES OF SYSTEM

Unique features of the proposed system to overcome the limitations of existing system for sepsis detection are given below:-

- To provide sepsis detection within time without any delays.
- More accurate and efficient.
- Reliable.
- Fast and User friendly.
- Less wastage of resources.
- Time saving.
- Lesser possibility of false results.
- No man power required for detection.

CHAPTER 4

REQUIREMENT ANALYSIS AND SYSTEM SPECIFICATION

4.1 FEASIBILITY STUDY

The feasibility study for sepsis detection system covers three main aspects that are technical, economical and operational.

Technical feasibility

The system's technical feasibility is robust, utilizing of advance and modern technologies such as machine learning for training the model and HTML, CSS for web interface.

- **Machine learning**
  The process of teaching computers to learn patterns from data and make predictions without being explicitly programmed.
- **HTML**
  Hyper text markup language which is used to create the basic structure or body of web pages.
- **CSS**
  Cascading style sheets is a styling language which is used to describe the presentation

`

how the document will appear which is written either in HTML or XML.

These technologies are widely used in different sectors as per the requirements. These technologies are well supported, ensuring stability, scalability and reliability of the developed systems. Also there are various advancements going on these technologies by the help of which we can easily upgrade our system's overall performance. Accuracy and reliability of systems that are developed using machine learning methods are always high.

ECONOMICAL

The project's economical feasibility is promising as it aims to detect sepsis efficiently within time without resulting in any delays which comes along with various benefits such as reduced hospital stays, lower treatment costs, less wastage of hospital resources, better staff utilization and ultimately improved patient outcomes and leading to less operational costs for hospitals.

OPEARTIONAL

The operational feasibility of system is high as it is easy to use there's no as such need for giving training to get used to with the system. It can be easily accepted by healthcare professionals because of its ease of use and it can be integrated with health care systems after testing it in collaboration with doctors in real world scenarios.

4.2 SOFTWARE REQUIREMENT SPECIFICATION

4.2.1 DATA REQUIREMENT

Patient Data
The system only requires the patient data for whom we want to predict sepsis. There is no need of any user sign in or authentication only patient data with necessary values is required. The patient data will be needed to be provided to the system by doctors for predicting the onset of sepsis at an early diagnostic stage to improve the patient outcomes.

The patient data that will be provided by the doctor will typically include the following values :-
- Respiratory rate, Temperature, Mean arterial Pressure and other Vital Signs.
- Platelet Count, Glucose, Calcium etc. are Laboratory Values.
- Age, Gender, Time in ICU, Hospital Admit time etc. are considered as Demographics.

Patient data including these values will be provided to the system to effectively predict the onset of sepsis timely without any delays so that required actions can be taken immediately.

4.2.2FUNCTIONAL REQUIREMENT

Functional requirement analysis entails a thorough examination, analysis, and description of software requirements and hardware requirements in order to meet actual and also necessary criteria in order to resolve an issue. Analyzing functional Requirements includes a number of processes. We can say that the Functional Requirements for our project sepsis detection system will typically include the specific features and functionalities that the system must possess to effectively detect the onset of sepsis in patients which can include:

`

- Design a user-friendly interface that allows healthcare professionals to easily access and interpret patient data and alerts generated by the system.
- Easily upload the patient data and process it for predicting sepsis.
- Easy to understand so that doctors can easily adapt to the system.
- Use of efficient algorithms for predicting sepsis.
- Optimize the interface for various devices to provide seamless experience.

Software Requirement
- Machine learning.
- Python and libraries.
- Google colab.

Hardware Requirement
- Modern Operating System (windows 7 or 10/Mac OS X 10.11 or higher).
- x86 64-bit CPU.
- Disk Space - 4GB SSD.
- RAM/Main Memory - 4GB DDR4 3200Mhz.

4.2.3 PERFOMRANCE REQUIREMENT

Performance requirements for sepsis detection project will include the requirements that are essential for ensuring that the system operates efficiently and effectively for predicting the onset of sepsis in patients. These requirements will include the expectations related to the system's performance metrics. Here are some key performance requirements:

- **Response Time**
  The system should have fast response time for processing the provided patient data and displaying the prediction results without any delays.
- **Accuracy**
  The system should be able to predict the onset of sepsis with accuracy with lesser possibility of false results.
- **Scalability**
  The system must be able to handle increased user load without compromising with the performance or response time in displaying results.

- **Training time**
  The training time should be less for training the healthcare professionals to get easily adapt to the system.
- **Reliability**
  The system should be robust enough to operate effectively without many frequent disruptions.

4.2.4 MAINTAINABILITY REQUIREMENT

Maintainability requirements for sepsis detection system specify the criteria that are related to the ease of maintaining, updating, and supporting the system over time. These requirements are crucial for ensuring that the system remains effective, reliable, and adaptable throughout its lifecycle without any negative impacts on its performance. Here are some key maintainability requirements:
`

- **Modularity**

  Design the system with modular architecture to facilitate easy modification or updates to individual modules or components of the system without affecting the entire system. This promotes flexibility for changes and simplifies maintenance efforts.

- **Documentation**

  Ensure thorough documentation covering various components such as the system architecture, codebase, algorithms, and user manuals. Provide clear and concise comments within the code wherever needed to explain the logic and functionality. Clear documentation aids to provide better understanding of the system's functionality, issues that can occur, and also easy on boarding for users.

- **Coding Standards**

  Adhere to a consistent coding style to maintain readability and consistency across the code base. Applying coding standards and following best practices allow to maintain a clean and well-structured codebase which is easy to understand, debug or modify. Conduct regular code reviews to identify and address potential issues at an early stage.

- **Continuous Improvement**

  There is always a scope of making improvements in the system due to fast moving technology and requirements so we need to establish processes such as collecting user feedback, monitoring system performance, and incorporating enhancements or optimizations iteratively. By using these feedback there is always a scope of improvement to adapt to evolving requirements and technological advancements.

- **Testing**

  Implement testing practices to ensure robustness and reliability of the system. Define test cases covering various scenarios, including edge cases and error conditions. Design the system for testability, making it easy to isolate and test individual components of the system making it effective at an early stage by testing it for multiple scenarios.

By incorporating above mentioned maintainability requirements into our project plan and development process, we can ensure the long-term reliability and sustainability of the sepsis detection system, and ultimately maximizing its worth for both to healthcare providers and patients.

4.3 SDLC MODEL

Waterfall model will be adopted for the development of sepsis detection system because it is characterized as a sequential approach towards software development. It is widely used in situations where the project requirements are well defined and project goals are clear. In our case the project goal is to build a model for detecting sepsis at an early stage effectively so we can adopt for waterfall model.

`
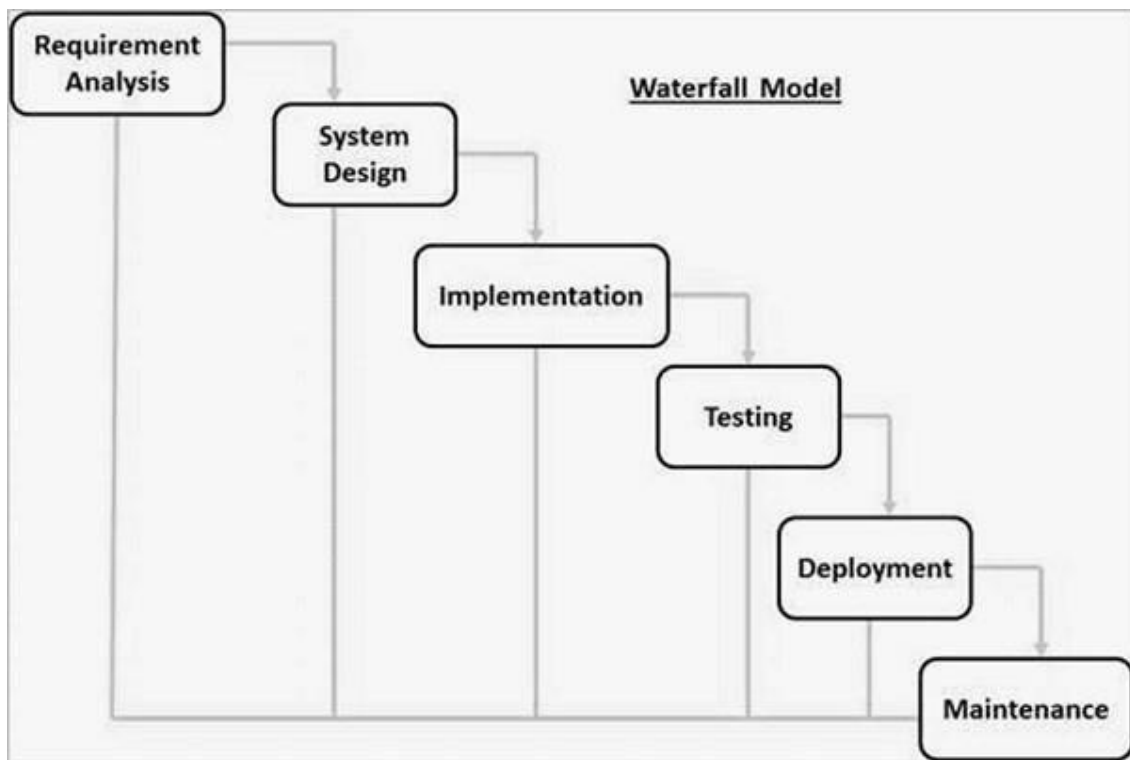
Fig-2 SDLC Model

4.3.1 SYSTEM DESIGN

This chapter consists of the design of the software Life Cycle model diagrams such as data flow diagrams, use case diagram, workflow about how the system will proceed etc with their detailed explanation. Design is about choosing the right architecture and solutions that are appropriate to the problem. Once the requirements are finalized the system architecture will be designed using machine learning techniques. Keeping the requirements in mind the system specifications are translated into a software representation. In this phase the designer emphasizes on:- algorithms that are to be used, data structure, software architecture etc.
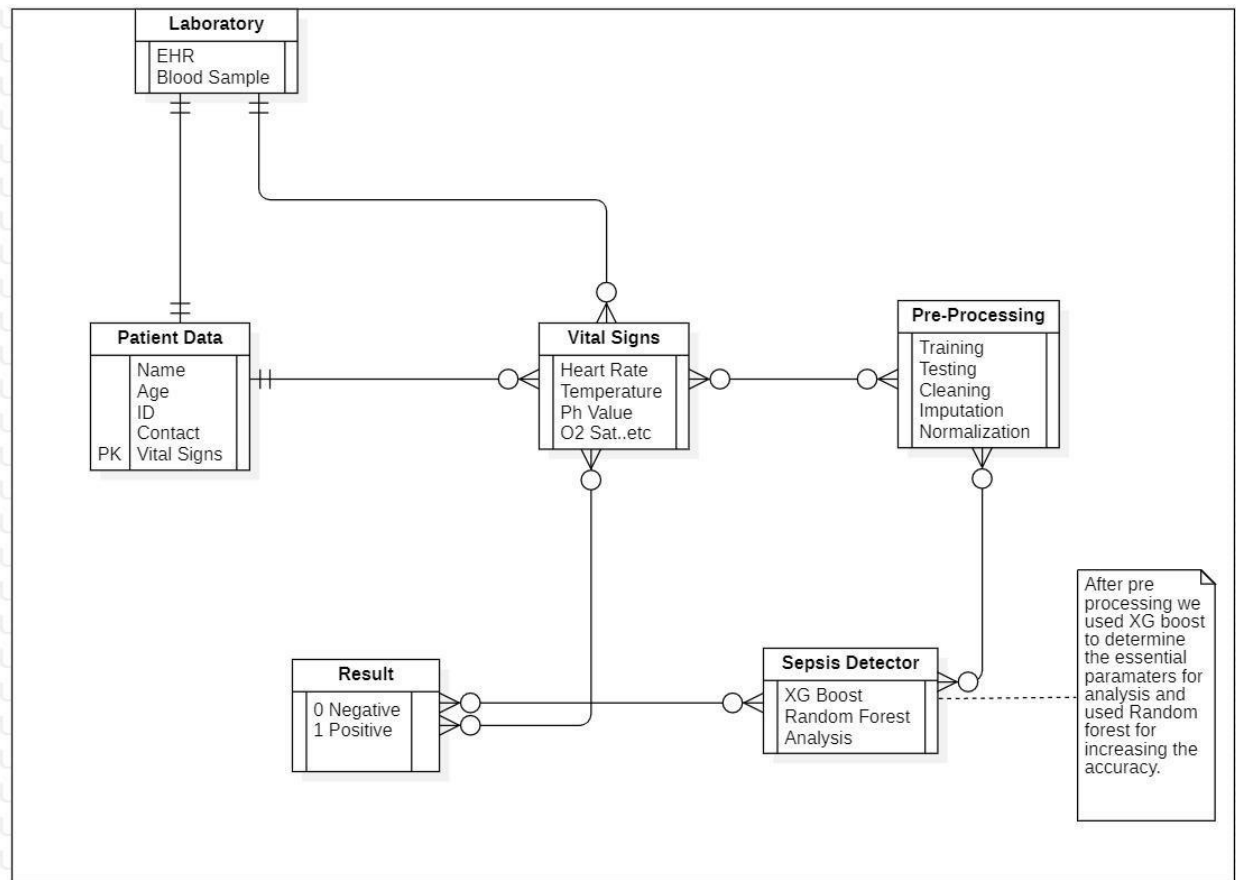
`
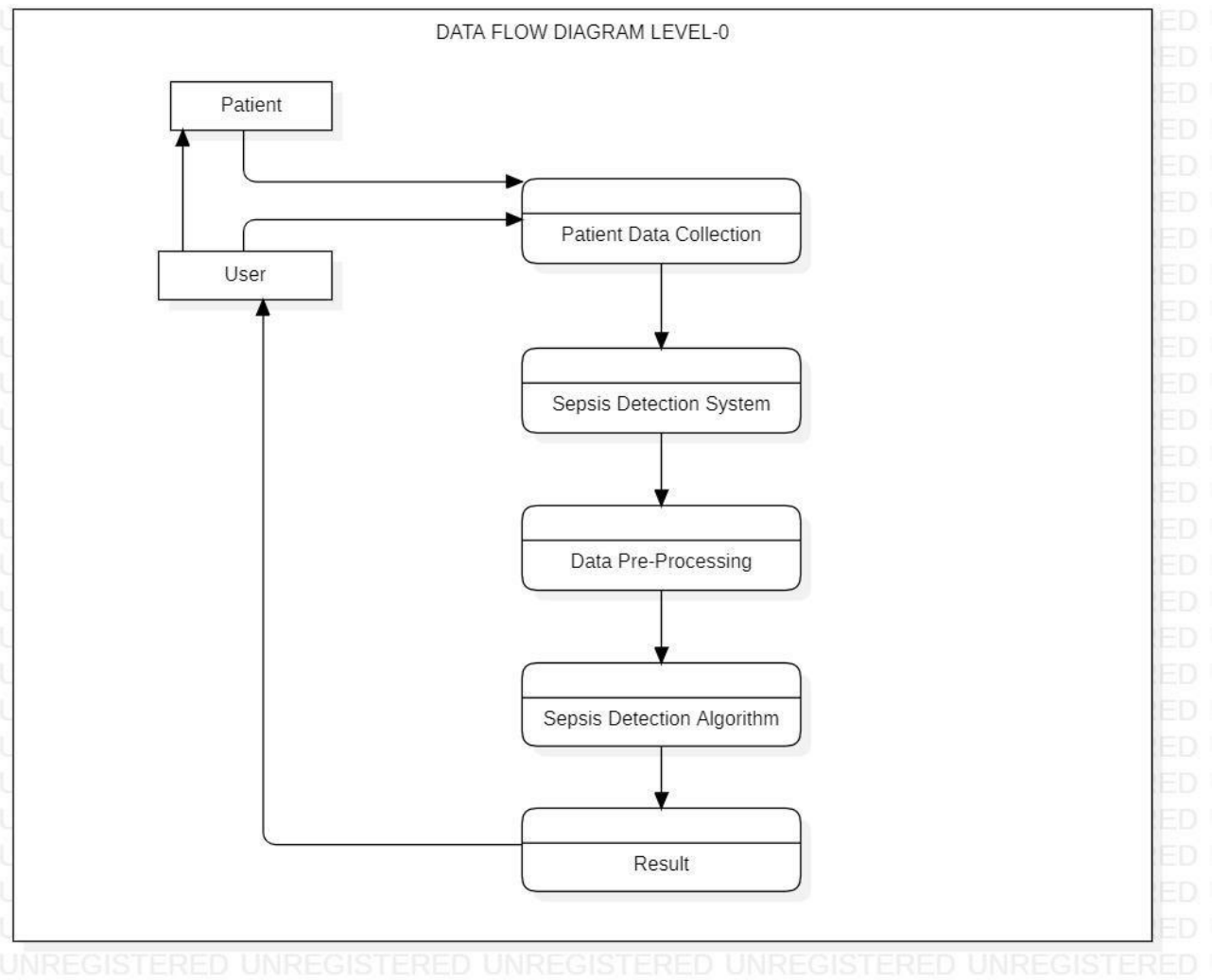
Fig -3 DFD Level 0
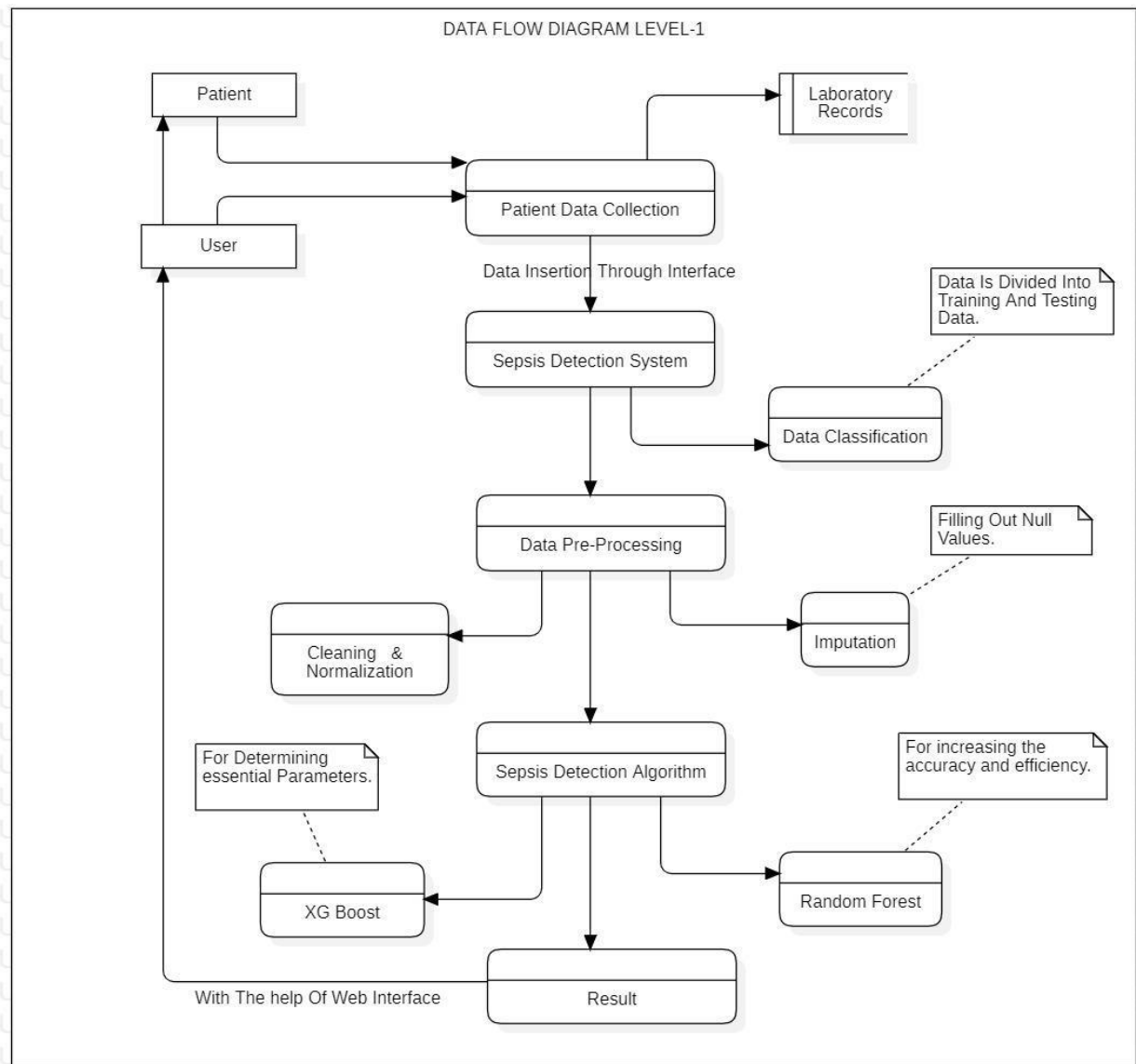
Fig -4 DFD Level 1
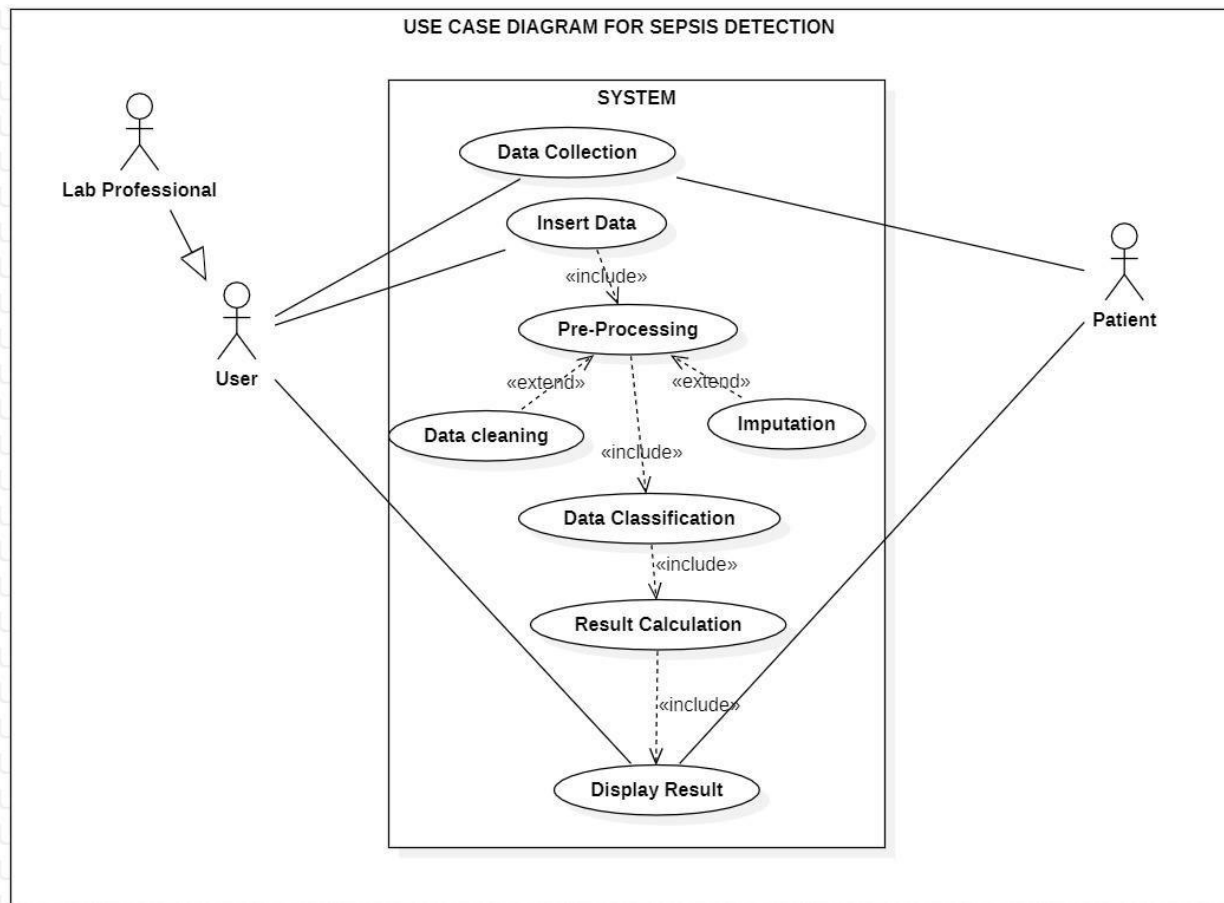
Fig -5 Use Case Diagram

Fig -6 ER Diagram
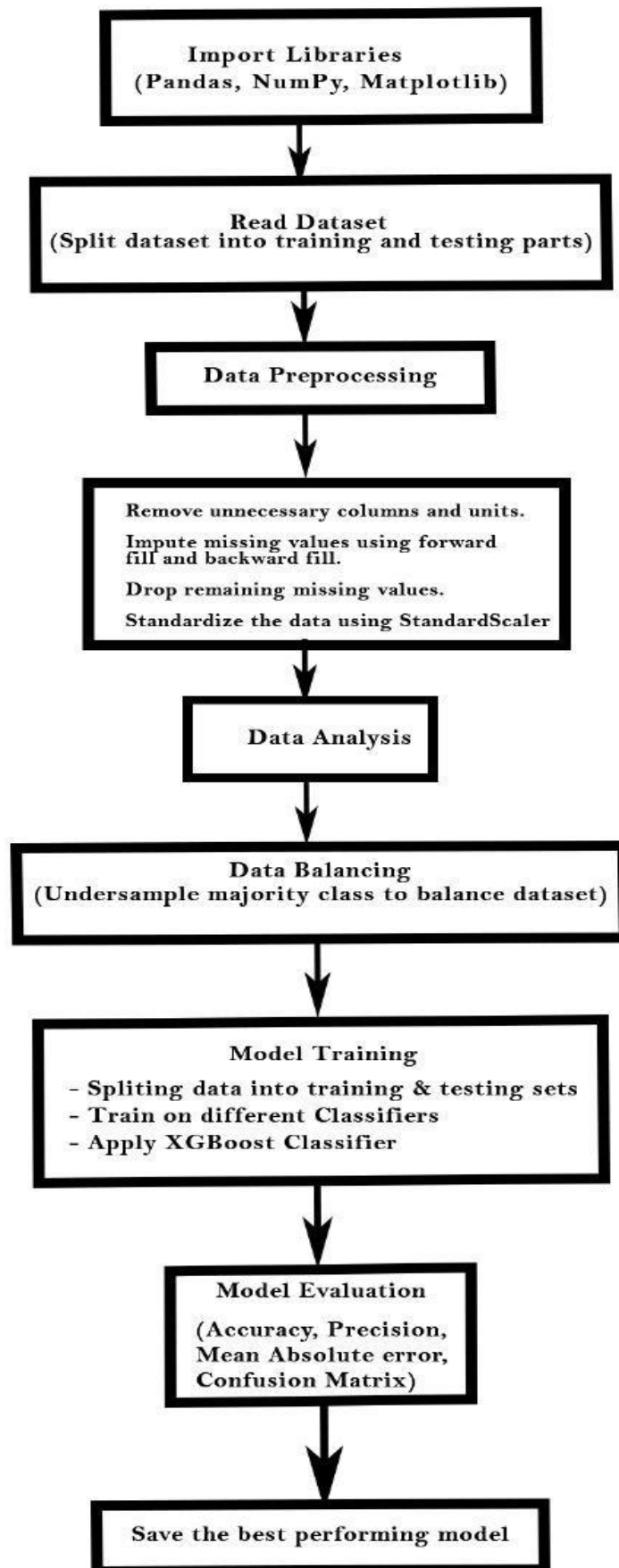
Fig -7 Workflow Diagram

Explanation:-

The above flow chart represents the system architecture how the system will work step by step for predicting sepsis. The steps involve:

- Collect our dataset
- Refine it and handle if any errors
- Data preprocessing
- Define the model
- Train the model using different algorithms
- Check model on testing data
- Model evaluation
- Choose the best performing model
- If desirable then use this model to predict sepsis on real values/ scenario

`

CHAPTER 5

IMPLEMENTATION

5.1 Introduction to Languages, and Technologies

The implementation phase involves the actual coding or programming of the software. The output of this phase is typically the library, executable, user manuals and additional software documentation.

FRONTEND

For building the web interface which will be visible to the users and the healthcare professionals will connect to it for predicting sepsis at an early stage for better diagnosis will be built using HTML and CSS. HTML and CSS both are part of building the front end part of any website with which the user interacts.

HTML

HTML refers to (Hypertext Markup Language) is the standard language (for building the front end part of any website) which is used to create and design the basic structure or body of web pages. It uses simple tags to structure content, such as headings, paragraphs, lists, and links. These tags allow users to view text, images, videos, and other media on the internet. With HTML, we can create the layout and format of a web page, making it easy to organize and present the desired information.

CSS

CSS refers to (Cascading Style Sheets) is a language which is used for styling and formatting web pages. It allows us to control the appearance of HTML elements, such as text, colors, fonts, spacing, background and layout. By using CSS, we can enhance the visual representation of a web page, making it more appealing and easier to read. CSS can be specified directly within HTML documents or in separate CSS files, which leads to provide flexibility and consistency in design across multiple pages of a website.

BACKEND

For building model that will predict sepsis we will use machine learning methods and algorithms because of its accuracy and reliability. This will process the provided patient data and will help healthcare professionals to predict the onset of sepsis within minimal time without any delays.
`

MACHINE LEARNING

Machine learning is an application of artificial intelligence (AI) that gives systems the ability to automatically learn and evolve from experience without being specially programmed by the programmer. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The main aim of machine learning is to allow computers to learn automatically and adjust their actions to improve the accuracy and usefulness of the program, without any human intervention or assistance In simpler terms, it's about creating algorithms that can learn from and make predictions or decisions based on data. This enables machines to perform tasks such as recognizing faces in photos, understanding spoken language, recommending products, and much more. In essence, machine learning helps computers learn from experience and improve their performance over time. Because of various reasons machine learning is an effective choice for making decisions or predictions not only in health care sector but in various industries according to their requirements which is done by observations , finding patterns on large data with accuracy.

PYTHON

Python is one of the most popular programming languages used for machine learning due to its simplicity, versatility, and extensive libraries. With Python, we can easily implement various machine learning algorithms, preprocess the data, visualize results, and deploy models. Some of the key libraries for machine learning in Python include:

Pandas

 In computer programming, pandas is a data manipulation and analysis software package designed for the Python programming language. It includes data structures and methods for manipulating numerical tables and time series, in particular.

Seaborn

Seaborn is an amazing visualization library used for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. Seaborn aims to make visualization the central part of exploring and understanding data.

Sklearn

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It

`

provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. A user-friendly library for classical machine learning algorithms such as regression, classification, clustering. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Matplotlib

Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension Numpy. As such, it allows us to create insightful plots and graphs to analyze data and visualize model results. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications.

In summary, we can conclude that Python provides a rich ecosystem of libraries and tools for machine learning, making it an ideal choice for both beginners and experienced practitioners in the field. Its simplicity, readability make it a versatile language for developing machine learning solutions across various domains.

TOOLS

GOOGLE COLAB

Google Colab or Google Colaboratory is a free cloud-based platform which is provided by Google that allows users to write and execute Python code in a browser-based environment, without requiring any setup or installation. It allows you and your team to collaborate on projects in the same way that you do with Google Docs. Many common machine learning libraries are supported by Colab and can be quickly loaded into your notebook.

Google Colab also provides integration with Google Drive, allowing users to save and share their Colab notebooks directly in their Google Drive accounts. This feature makes it convenient for collaboration and sharing of code and results with others.

It is a valuable tool for working on Python-based projects, particularly in the fields of machine learning, data science, and artificial intelligence.

CHAPTER 6 TESTING AND MAINTENANCE

6.1 Testing

Comprehensive testing will be conducted after the development of system to ensure functionality, performance and accuracy according to the requirements.
`

Manual Testing

Test Case Id	Input	Expected O/P	Actual O/P	Status
1	Raw data is loaded	Data set is cleaned	Data set is cleaned	Pass
2	File format .txt	Predict results	Predicted results	Pass
3	File format .png	Error	Error	Pass
4	File format .jpg	Error	Error	Pass
5	File format .word	Error	Error	Pass
6	Data of patient with sepsis	Predict patient have sepsis	Predict patient have Sepsis	Pass
7	Data of patient with No Sepsis	Predict Patient doesn't have Sepsis	Predict patient doesn't have sepsis	Pass
8	Correct URL to load web interface	Web interface will be loaded	Web interface will be loaded	Pass
9	Incorrect URL for interface	Error	Error	Pass

Decision Table

In the given decision table C1, C2 refers to the conditions and A1, A2 refers to the respective actions based on the conditions. This table shows that only txt format file will be accepted to predict sepsis and other formats will display error.

	C1- txt format	C2 - jpeg format	C3 - png format	C4 - Blank txt file
A1- Accepted and predict sepsis.	True			
A2-Bad request error		True	True	
A3-Not enough values				True

Screenshot

`

CHAPTER 7 (RESULTS AND DISCUSSIONS)
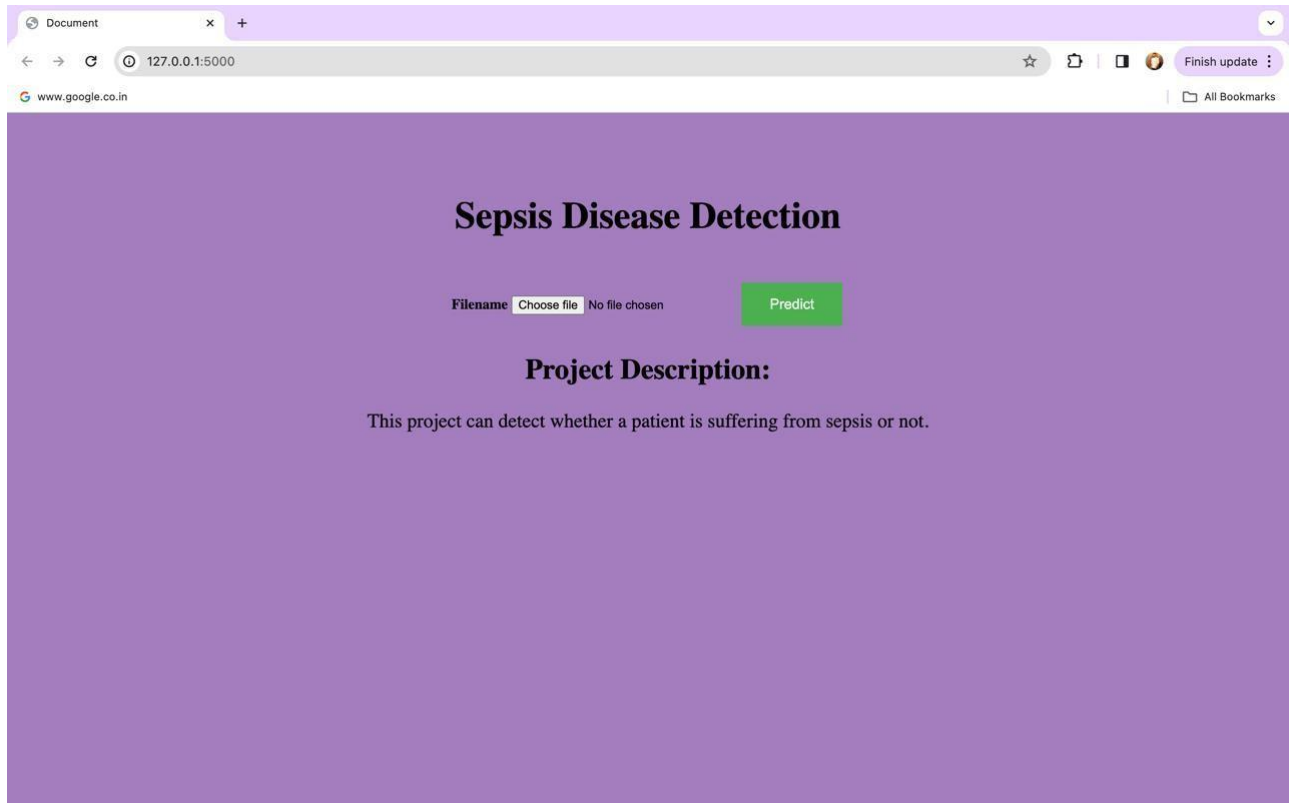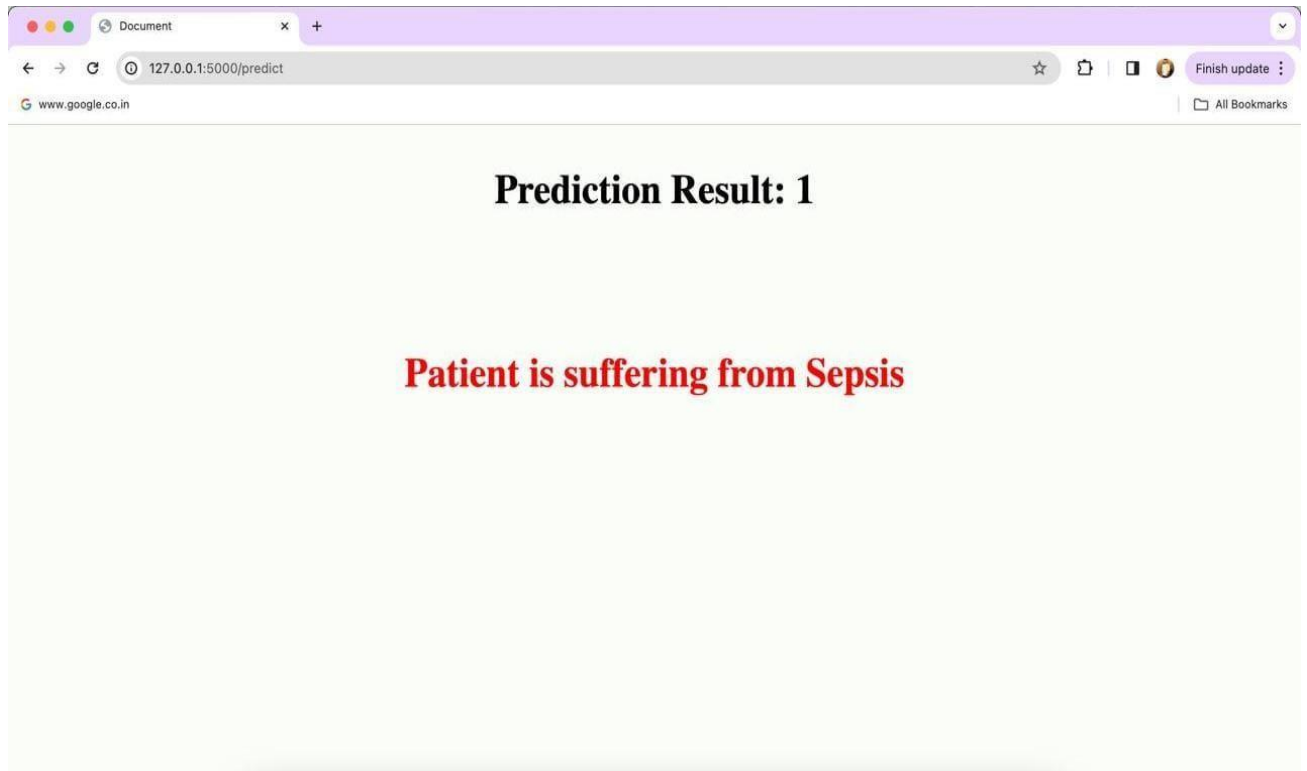
7.1 User Interface Representation

FIG 7.1.1

OUTPUT SCREENSHOT

FIG 7.1.2

`

FIG 7.1.2

7.2 Brief Description of Various Modules

This chapter presents the brief description part for all the modules that are involved in our project.

7.2.1 Pre Processing

For the raw dataset obtained on which we will train our model there are lots of missing values or null values for which imputation is done to fill the missing values. In the further modified train dataset, we removed the feature columns that are having null values greater than 25%. One hot encoding is implemented for gender column to convert categorical data into numerical data or binary format which machine learning models can easily understand, Standard normalization for the remaining columns is implemented and NA values are dropped for the dataset to improve the performance and stability of applied machine learning algorithms. The number of rows with sepsis labels as 0 and 1 are not equal, so undersampling of the dataset is done in order to balance the dataset.

7.2.2 Feature Selection

For the pre-processed dataset, we generated the correlation matrix and removed the features that are having high correlation and drop those feature columns from the train dataset. This helps to simplify our dataset and removes unnecessary redundancy, making it cleaner and potentially improving the performance of our machine learning model. This process is often done to optimize

`

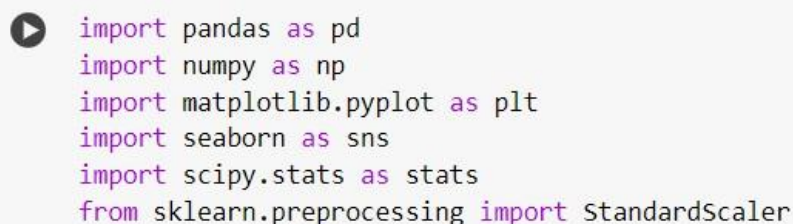the preprocessed dataset for better model performance and accuracy.

7.2.3 Model Training and Evaluation

After preprocessing we train our dataset on different machine learning algorithms such as naïve bayes, random forest, knn etc and measuring the performance of the models based on their accuracy and identified the best classifiers among the trained models. Based on the evaluation results, provide insights into which machine learning algorithm demonstrates the highest level of accuracy for predicting sepsis on the training dataset. After performing evaluation on different models the best classifier among the considered classifiers is Random Forest Classifier with an accuracy of almost 97 percent.

This above given brief description outlines the simple methodology used for comparing the accuracy of different machine learning algorithms on a training dataset for predicting the onset of sepsis at an early stage and considering the best classifier for predicting sepsis.

7.3 Snapshots of System

7.3.1 Importing all the libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
from sklearn.preprocessing import StandardScaler
```

Fig.7.3.1. Imported Libraries Code Snippet

Here we are showing that we are importing all the necessary files and libraries which are required for our ml model.

7.3.2 Data Fields

The patient raw dataset includes various values which will undergo preprocessing and further for training on models such as:

- Respiratory rate, Temperature, Mean arterial Pressure and other vital signs.
- Platelet Count, Glucose, Calcium and some other laboratory values.
- Age, Gender, Time in ICU, Hospital Admit time etc. are considered as Demographics.
- 0 (Non-sepsis) and 1 (Sepsis) is the label for identification.

`

	Unnamed: 0	Hour	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	...	Fibrinogen	Platelets	Age	Gender	Unit1	Unit2	HospAdmTime	ICULOS	SepsisLabel	Patient_ID
0	0	0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	1	0	17072
1	1	1	65.0	100.0	NaN	NaN	72.0	NaN	16.5	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	2	0	17072
2	2	2	78.0	100.0	NaN	NaN	42.5	NaN	NaN	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	3	0	17072
3	3	3	73.0	100.0	NaN	NaN	NaN	NaN	17.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	4	0	17072
4	4	4	70.0	100.0	NaN	129.0	74.0	69.0	14.0	NaN	...	NaN	330.0	68.54	0	NaN	NaN	-0.02	5	0	17072
5	5	5	62.0	100.0	NaN	124.0	85.0	61.0	14.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	6	0	17072
6	6	6	61.0	100.0	NaN	101.0	75.0	58.0	14.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	7	0	17072
7	7	7	68.0	100.0	35.78	142.0	93.5	78.0	16.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	8	0	17072
8	8	8	71.0	100.0	NaN	121.0	74.0	91.0	14.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	9	0	17072
9	9	9	69.0	100.0	NaN	120.0	79.0	98.0	14.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	10	0	17072
10	10	10	75.0	100.0	NaN	146.0	93.0	67.0	14.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	11	0	17072
11	11	11	84.0	100.0	36.39	128.0	80.0	60.0	14.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	12	0	17072
12	12	12	85.0	100.0	NaN	124.0	79.0	59.0	14.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	13	0	17072
13	13	13	85.0	100.0	NaN	141.0	95.0	69.0	14.0	NaN	...	NaN	303.0	68.54	0	NaN	NaN	-0.02	14	0	17072
14	14	14	89.0	100.0	NaN	117.0	86.0	68.0	14.0	NaN	...	NaN	NaN	68.54	0	NaN	NaN	-0.02	15	0	17072

Fig.7.3.2. Data Set Snippet 1

```
df_train_impute.head()
```

	Hour	HR	O2Sat	Temp	MAP	Resp	BUN	Chloride	Creatinine	Glucose	Hct	Hgb	WBC	Platelets	Age	HospAdmTime	ICULOS	SepsisLabel	0	1
0	0	65.0	100.0	35.78	4.290459	16.5	3.178054	104.0	0.587787	5.087596	29.7	9.5	2.509599	5.802118	68.54	-0.02	1	0	1	0
1	1	65.0	100.0	35.78	4.290459	16.5	3.178054	104.0	0.587787	5.087596	29.7	9.5	2.509599	5.802118	68.54	-0.02	2	0	1	0
2	2	78.0	100.0	35.78	3.772761	17.0	3.178054	104.0	0.587787	5.087596	29.7	9.5	2.509599	5.802118	68.54	-0.02	3	0	1	0
3	3	73.0	100.0	35.78	4.317488	17.0	3.178054	104.0	0.587787	5.087596	29.7	9.5	2.509599	5.802118	68.54	-0.02	4	0	1	0
4	4	70.0	100.0	35.78	4.317488	14.0	3.178054	104.0	0.587787	5.087596	29.7	9.5	2.509599	5.802118	68.54	-0.02	5	0	1	0

Fig.7.3.3 Data Set Cleaned

7.3.3 Loading, Pre-processing and Splitting the data

Loading a dataset requires reading the data from a file or other source and storing it in memory in order to analyze and model it. CSV, JSON, and Excel are just a few of the formats in which datasets can be saved. Depending on the format of the dataset, various tools and libraries can be used to load the data. With utilities for reading CSV, Excel, and other formats, pandas is a popular package which is used for reading and editing datasets or data frames in Python.

After importing the dataset, pre-processing is frequently necessary to prepare the raw data for analysis and modeling. This aids in variable transformation, missing value management, and data formatting and cleaning. It is an important step in the data analysis process because it greatly affects the analysis's results and the model's performance and accuracy.

After preparing the data, it is usual to split the dataset into a training set and a test set. The training set is used to fit the model, and the test set is used to evaluate the model's performance. By segmenting the dataset in this way, data from the training set are included in the test set, enabling an unbiased evaluation of the model. In order to ensure that the training and test sets are accurate representations of the complete dataset, they must have been selected at random.

`

7.3.3 Applying Algorithms Screenshots

```python
# KNN Classifier
# KNN was tested on different values of k

from sklearn.neighbors import KNeighborsClassifier
# model = KNeighborsClassifier(n_neighbors=8)
# model = KNeighborsClassifier(n_neighbors=5)
model = KNeighborsClassifier(n_neighbors=10)
model.fit(X_train, y_train)
knn_predictions = model.predict(X_test)
evaluate_model(y_test,knn_predictions)
```

```
Accuracy: 0.8308799476611056
Precision: 0.7822892498066512
Recall: 0.6718698106941215
F1 Score: 0.7228872610326962
AUC-ROC: 0.790236853399009
Mean Absolute Error: 0.16912005233889435
Root Mean Squared Error: 0.4112420848343398
```

Fig.7.3.4. Applying KNN Code Snippet

```python
X = df_train_impute.drop('SepsisLabel', axis=1)
y = df_train_impute['SepsisLabel']
X.columns = X.columns.astype(str)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
```

```python
model = RandomForestClassifier(n_estimators=300, random_state=0)
model.fit(X_train, y_train)
rcf_predictions = model.predict(X_test)
```

```python
evaluate_model(y_test,rcf_predictions)
```

```
Accuracy: 0.9657616399520227
Precision: 0.9342995169082126
Recall: 0.9634672866157423
F1 Score: 0.9486592544146502
AUC-ROC: 0.9651752017494296
Mean Absolute Error: 0.03423836004797732
Root Mean Squared Error: 0.1850361047146673
```

Fig.7.3.5. Applying Random Forest Code Snippet

`

```
# Naive Bayes Classifier
# NBC performed worse than random forest on each and every aspect of the evaluation metrics

from sklearn.naive_bayes import GaussianNB
model = GaussianNB()
model.fit(X_train, y_train)
nbc_predictions = model.predict(X_test)
evaluate_model(y_test,nbc_predictions)
```

```
Accuracy: 0.7581506923999564
Precision: 0.7105682421667552
Recall: 0.4443706409830621
F1 Score: 0.5467919901920719
AUC-ROC: 0.6779483075045181
Mean Absolute Error: 0.2418493076000436
Root Mean Squared Error: 0.4917817682672301
```

Fig.7.3.6 Applying Naïve Bayes Code Snippet

In the above methodology we have compared different machine learning algorithms for predicting the onset of sepsis with accuracy by giving input parameters which include patient vital signs and other parameters and selecting the Random Forest Algorithm for sepsis prediction.

`

CHAPTER 8

CONCLUSION AND FUTURE SCOPE

8.1 CONCLUSION

In this study, we have explored the performance of various machine learning algorithms for predicting the onset of sepsis in minimal time without any delays using a training data set containing vital signs, laboratory results, and patient demographics. After thorough experimentation and evaluation, it was found that Random Forest exhibited the highest level of accuracy in predicting sepsis on the given data set, achieving an accuracy score of almost 97 percent. Four different categories of artificial intelligence algorithms were the subject of the investigation for predicting the onset of sepsis, and concluded that Random Forest shows the greatest accuracy with almost 97% among all other considered in this research.

This finding suggests that Random Forest may be a suitable choice for sepsis prediction among the other considered algorithms offering promising results in terms of accuracy. Furthermore, it is worth noting that the performance of machine learning algorithms may vary depending on the specific characteristics of the data-set and the preprocessing techniques applied. As we all know that there is always a need of enhancing the model with time therefore, continued exploration and refinement of the predictive models are always necessary to optimize their performance and robustness not only in clinical practice but in other sectors also.

Overall, this study contributes valuable insights into the application of machine learning techniques for sepsis prediction, ultimately improving patient outcomes through early detection and intervention strategies. Future research may focus on exploring ensemble methods, and validating and testing the predictive models in collaboration with healthcare professionals to further enhance the accuracy of sepsis prediction algorithms for getting it implemented in real world scenarios.

8.2 FUTURE SCOPE

In the future, it is anticipated that this trained model would be used in real world applications for predicting sepsis at an early stage. This interface will have a very high future scope because it can serve the emergent needs of doctors to predict sepsis accurately at an early stage. Collaborating with healthcare providers for real world testing and making changes accordingly before getting it deployed as website in hospitals to improve the overall accuracy and reliability of the system. In future, we would like to enhance the application by making it more customer specific and deploying this model in hospital websites and help the doctors to detect early signs of the disease and also by predicting which type of sepsis and how much the patient is affected by the disease.

Collaborate with medical professionals for more deep research, to ensure a better understanding of sepsis and other related information which is critical for them related to sepsis to make advancements in the accuracy and usability of build predictive models can help to get these systems deployed for real world application. Additionally, it is anticipated that in the future, this model will be tested with additional machine learning algorithms to determine an effective way to predict the results and that work will be done on the application's user interface to make it easier for end users to use and to do so with the least amount of training or prior knowledge.

`

CHAPTER 9

REFERENCES

[1] - Baldominos, A., Puello, A., Oˇgul, H., A¸suroˇglu, T., & Colomo-Palacios, R. (2020). "Predicting infections using computational intelligence–a systematic review." IEEE Access, 8, 31083-31102.

[2] - Al-Mualemi, B. Y., Lu, L. (2020). A deep learning-based sepsis estimation scheme. IEEE Access, 9, 5442-5452.

[3] - Feng, M., McSparron, J. I., Kien, D. T., Stone, D. J., Roberts, D. H., Schwartzstein, R. M., .. Celi, L. A. (2018). "Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database." Intensive Care Medicine, 44(6), 884-892.Jia, Z. (2009).

[4] - Kam, H. J., & Kim, H. Y. (2017). "Learning representations for the early detection of sepsis with deep neural networks." Computers in Biology and Medicine, 89, 248-255.

[5] - Wang, X., Wang, Z., Weng, J., Wen, C., Chen, H., & Wang, X. (2018). "A new effective machine learning framework for sepsis diagnosis." IEEE Access, 6, 48300-48310.

[6] - Zhao X, Shen W, Wang G. Early Prediction of Sepsis Based on Machine Learning Algorithm. Comput Intell Neurosci. 2021 Oct 12;2021:6522633. doi: 10.1155/2021/6522633. PMID: 34675971; PMCID: PMC8526252.

[7] - Wang Z, Yao B. Multi-Branching Temporal Convolutional Network for Sepsis Prediction. IEEE J Biomed Health Inform. 2022 Feb;26(2):876-887. doi: 10.1109/JBHI.2021.3092835. Epub 2022 Feb 4. PMID: 34181558.

[8] - M. M. Islam, T. Nasrin, B. A. Walther, C. C. Wu, H. C. Yang, and Y. C. Li, "Prediction of sepsis patients using machine learning approach: A meta-analysis," Comput Methods Programs Biomed, vol. 170, pp. 1-9, Mar. 2019. doi: 10.1016/j.cmpb.2018.12.027. Epub 2018 Dec 26. PMID: 30712598.

[9] - M. Fu, J. Yuan, M. Lu, P. Hong, and M. Zeng, "An ensemble machine learning model for the early detection of sepsis from clinical data," in 2019 Computing in Cardiology (CinC), IEEE, September 2019.

[10] - T. X. Ying and A. Abu-Samah, "Early Prediction of Sepsis for ICU Patients using Gradient Boosted Tree," 2022 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS), Shah Alam, Malaysia, 2022, pp. 78-83, doi: 10.1109/I2CACIS54679.2022.9815467.

[11] K. Kalaiselvi and M. Deepika, "Machine Learning for Healthcare Diagnostics," in Machine Learning with Health Care Perspective, V. Jain and J. Chatterjee, Eds. Springer, Cham, Learning and Analytics in Intelligent Systems, vol. 13, pp. [91-105], 2020. doi: 10.1007/978-3-030-40850-3_5.

[12] H. Habehh and S. Gohel, "Machine Learning in Healthcare," Curr. Genomics, vol. 22, no. 4, pp. 291-300, Dec. 2021.

[13] Q. An, S. Rahman, J. Zhou, and J.J. Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges," Sensors, vol. 23, no. 9, p. 4178, 2023.

`