Sepsis Prediction Using Random Forest Classifier

Major Project report submitted in partial fulfillment of the
Requirements for the Award of the Degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

By

A.L Anupama 198W1A0566
A.V.L Lahari 198W1A0567

Under the Guidance of

Mr. S. Babu M.Tech., (Ph.D.)

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
V.R SIDDHARTHA ENGINEERING COLLEGE
Autonomous and Approved by AICTE - Twice Accredited by NBA
Affiliated to Jawaharlal Nehru Technological University, Kakinada
Vijayawada - 520007

2023

V.R SIDDHARTHA ENGINEERING COLLEGE

(Autonomous)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the major project report entitled **"Sepsis Prediction Using Random Forest Classifier"** being submitted by

A.L Anupama	198W1A0566
A.V.L Lahari	198W1A0567

in partial fulfilment for the award of the Degree of Bachelor of Technology in Computer Science and Engineering to the Jawaharlal Nehru Technological University, Kakinada is a record of bonafided work carried out during the period from 2022 - 2023 under my guidance and supervision.

Mr. S. Babu, M.Tech., (Ph.D.)
Assistant Professor & Guide

Dr. D. Rajeswara Rao, M.Tech., Ph.D.
Head of the Department

DECLARATION

We hereby declare that the dissertation entitled "Sepsis Prediction Using Random Forest Classifier" submitted for the B.Tech Degree is our original work and the dissertation has not formed the basis for the award of any degree, associate ship, fellowship or any other similar titles.

Place: Vijayawada A.L Anupama (198W1A0566)

Date: A.V.L Lahari (198W1A0567)

ACKNOWLEDGEMENT

We would like to thank **Dr. A. V. Ratna Prasad**, Principal of Velagapudi Ramakrishna Siddhartha Engineering College for the facilities provided during the course of Major Project.

We have been bestowed with the privilege of thanking **Dr. D. Rajeswara Rao**, Professor and Head of the Department for his moral and material support.

We would like to express our deep gratitude to our guide **Mr. S. Babu**, Assistant Professor for his persisting encouragement, everlasting patience and keen interest in discussion and for his numerous suggestions which we had at every phase of this project.

We owe our acknowledgements to an equally long list of people who helped us in major project work.

Place: Vijayawada

Date:

Abstract

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 [1]. Sepsis occurs when body's response to the chemicals is out of balance, triggering changes that can damage multiple organ substances. For the patient the practicality of detecting sepsis disease occurrence in development is an important factor in the result. The primary goal is to build models using Multi Layer Perceptron classifier, Adaboost classifier, Guassian Naïve bayes, Linear discriminant analysis, Gradient Boosting Classifier, Random Forest classifier to find out the best classifier with best performance measures and detect the sepsis disease in minimal time. Our secondary goal is to build and design a user-friendly web application. In this project, PhysioNet Challenge data is used to identify the best classifier. The proposed system comprises four stages. They are Pre-Processing, Feature Selection, Model Training and Model Evaluation. In this way, models are built using different classifiers to find out the best classifier with high-performance measures. Among the considered classifiers, Random Forest Classifier has better performance measures. Developed a Graphical User Interface using Flask to predict sepsis disease.

Keywords: Sepsis, Multi Layer Perceptron, Adaboost, Guassian Naïve bayes, Linear discriminant analysis, Gradient Boosting, Random Forest.

Table of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

This chapter contains basic concepts required for the project, problem statement, motivation, objectives, scope and advantages of the project.

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. Generally, the body releases natural synthetics into the circulation system in order to counterbalance the infection which is inside. Sepsis occurs when the body's response to these chemicals is out of balance, this can damage many organ systems. Sepsis is caused by infection and can happen to anyone. 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. This disease is a major health concern for the public in terms of morbidity, health care expenses and mortality. Detecting at early stages, with antibiotic treatment the outcomes can be improved [1]. Though many professional care societies have proposed new methods in recognising sepsis, the central requirement for early identification and treatment remains neglected. It can be treated if it can be recognised at early stages. Several examinations have demonstrated that delays in finding and treatment of sepsis can prompt high death rates. Our main motto is to detect sepsis as soon as the patient visits the emergency department for the treatment.

1.1 Basic Concepts

MLP Classifier

MLP Classifier also known as Multi-layer Perceptron classifier [1] which itself suggests a Neural Network. MLP Classifier relies on an elemental Neural Network to perform the classification task. It comes under ANN. The phrase MLP is used ineptly, sometimes roughly to refer any feedforward ANN, occasionally strictly referring to networks consisting multiple layers of perceptrons (with threshold activation) Multilayer perceptrons now and then are vernacularly referred as "vanilla" neural networks, notably if they contain a single hidden layer, avoiding long time-taking lab results. It is very flexible and can be used generally to learn a mapping from inputs to outputs. The model that is being built using MLP Clas-

sifier, the data which is obtained after preprocessing is given to the model and the pre-processed data is divided such that eighty percent for training the model and twenty percent used for testing the trained model [13]. With this MLP classifier we could achieve an accuracy of 94%, with a total of six layers in which one is input layer, four layers are considered as hidden layer and finally the last layer is the output layer, tanh as activation function and, max iterations up to 5000.

AdaBoost Classifier

AdaBoost, also called Adaptive Boosting, is a technique in Machine Learning used as an Ensemble Method. The most common estimator used with AdaBoost is decision trees with one level which means Decision trees with only 1 split. These trees are also called Decision Stumps. What this algorithm does is that it builds a model and gives equal weights to all the data points. It then assigns higher weights to points that are wrongly classified. Now all the points with higher weights are given more importance in the next model. It will keep training models until and unless a lower error is received.

Gradient Boosting classifier

The main idea behind this algorithm is to build models sequentially and these subsequent models try to reduce the errors of the previous model. But how do we do that? How do we reduce the error? This is done by building a new model on the errors or residuals of the previous model [6]. When the target column is continuous, we use Gradient Boosting Regressor whereas when it is a classification problem, we use Gradient Boosting Classifier. The only difference between the two is the "Loss function". The objective here is to minimize this loss function by adding weak learners using gradient descent. Since it is based on loss function hence for regression problems, we'll have different loss functions like Mean squared error (MSE).

Guassian Naive Bayes

Naïve Bayes is a probabilistic machine learning algorithm used for many classification functions and is based on the Bayes theorem [7]. Gaussian Naïve Bayes is the extension of naïve Bayes. While other functions are used to estimate data distribution, Gaussian or normal distribution is the simplest to implement as you will need to calculate the mean and standard deviation for the training data.Naïve Bayes is a probabilistic machine learning algorithm used for many classification functions and is based on the Bayes theorem.

Linear Discriminant Analysis

Logistic Regression is one of the most popular linear classification models that perform well for binary classification but falls short in the case of multiple classification problems with well-separated classes. While Linear Discriminant Analysis(LDA) handles these quite efficiently. LDA can also be used in data preprocessing to reduce the number of features just as PCA which reduces the computing cost significantly.

Random Forest Classifier

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 [14]. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks.

Sepsis Symptoms

Symptoms include:

- Fever: temperature above 38^{o}C or even the temperature below normal i.e., 36^{o}C.
- Heart rate: greater than 90 beats/min.
- Breathing: higher than 20 breaths/min.

Causes of Sepsis

Any infection can trigger sepsis, but the following types of infections are more likely to cause sepsis:

- Pneumonia: It is a type of infection that attacks the lungs one or both lungs can be affected by bacteria, fungi, and viruses from outside attack lungs. This causes inflammation in the air sacs called alveoli in the lungs, the bacteria or virus or fungi fills this with fluid which makes breathing difficult.
- Abdominal infection: It surrounds a number of infectious processes, including peritonitis, cholecystitis, diverticulitis, pancreatitis, and cholangitis. With help of Empirical treatment, they can identify whether the infection is through community or healthcare-acquired, the organs which are infected, and to check if the infection is complex or simple.

- Kidney Infection: It generally results from an infection in the urinary tract that spreads to 1 or both the kidneys, this can be chronic or sudden. If they are not treated at early stages they can be life-threatening.
- Bloodstream Infection: It is an infection that occurs when bacteria are in the circulatory system. It generally describes bacteraemia or sepsis. Sepsis is a serious, potentially fatal infection. This infection can cause sepsis to grow rapidly. Brief diagnosis and treatment are basic for treating this infection at the early stages.

Existing Systems

- LSTM: Long Short-Term Memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems.
- RNN: Recurrent Neural Networks (RNN) model is designed to recognize the sequential characteristics of data and thereafter using the patterns to predict the coming scenario.
- RBFNN: Radial Basis Function Neural Networks (RBFNN) is an ANN that uses radial basis functions as activation functions.

Drawbacks of Existing Systems

- Gradient vanishing and exploding problems. Training an LSTM is a very difficult task. It cannot process very long sequences.
- It cannot process very long sequences. Due to its recurrent nature, the computation is slow of an RNN model and only gives 82% accuracy.
- Classification will take more time with only an accuracy of 87% with an RBFNN model.

1.2 Problem Statement

Sepsis is a potentially life-threatening condition that occurs when the body's response to an infection and damages its own tissues. By detecting sepsis at early stages, helps in saving the lives of people. For the patient the practicality of detecting sepsis disease occurrence in development is an important factor in the result. Late prediction of sepsis is potentially life-threatening, and also consumes heavy hospital resources. By using Machine Learning models predicting sepsis consumes limited resources and can assume the risk of prediction to be minimal but revolutionary.

1.3 Objectives

The objectives of the project are:

- To identify the best classifier among Multi Layer Perceptron, Ada-Boost, Gradient Boosting, Gaussian Naive, Linear Discriminant classifier, Random Forest for prediction of sepsis at early stages.
- To design and develop a website using Flask and integrate with the model.

1.4 Scope

The scope of the project is:

- The patient should undergo some medical tests to predict whether he has affected.
- The patient got how much affected and which type of sepsis disease was he affected by is not shown. with sepsis disease.
- The project is only useful for doctors who has knowledge on the disease and its treatment.

1.5 Advantages

The advantages of the project are:

- Huge amount of data can be analysed in a minimal time.
- No man power is required for detection.
- Lesser possibility of false results.
- Time efficient.

Chapter 2

LITERATURE REVIEW

This chapter contains the list of research papers that we have studied under literature survey. We focused on the approaches for maintaining accuracy in these papers. Our study included the techniques used for developing and training the model.

2.1 Dynamic Sepsis Prediction for Intensive Care Unit Patients Using XGBoost-Based Model With Novel Time-Dependent Features [1]

This paper presented a sepsis prediction model based on XGBoost framework. The model used time-dependent feature construction method and statistical count construction method to construct some new features, which have more superior characteristics to express sepsis and greatly improve the model accuracy. A customized objective function for training XGBoost-based model was proposed also.

Advantages:

- The proposed model has the best performance, with AUROC improved by 5.4% on the MIMIC-III dataset and 2.1% on PhysioNet Challenge 2019 dataset.
- It has strong robustness and a wide range of application scenarios.
- The proposed model can be scaled for use in different healthcare settings and applications.
- The customized objective function used for training the XGBoost-based model helped in optimizing model performance and accuracy.

Disadvantages:

- The lower ranking features may have little effects on the results.
- Filling the data with linear interpolation could achieve better model performance

2.2 Early Prediction of Sepsis Based on Machine Learning Algorithm [2]

This study includes machine learning algorithms XGBoost and LightGBM are applied to construct two processing methods: mean processing method and feature generation method, aiming to predict early sepsis 6 hours in advance. The feature generation methods are constructed by combining different features, including statistical strength features, window features, and medical features. Miceforest multiple interpolation method is applied to tackle large missing data problems. Results show that the feature generation method outperforms the mean processing method. XGBoost and LightGBM algorithms are in good prediction performance.

When the amount of data and the number of features increase sharply to large scale, LightGBM has not only a fast iteration speed in training but also a better predictive ability than XGBoost. XGBoost has a stronger generalization ability when the amount of data and the number of features is relatively small. When the amount of data and the number of features increase sharply to a large scale, LightGBM has not only a fast iteration speed in training but also a better predictive ability than XGBoost. When the amount of data and the number of features increase sharply to a large scale, LightGBM has not only a fast iteration speed in training but also a better predictive ability than XGBoost.

Advantages:

- XGBoost has a stronger generalization ability when the amount of data and the number of features is relatively small.
- When the amount of data and the number of features increase sharply to a large scale, LightGBM has not only a fast iteration speed in training but also a better predictive ability than XGBoost.

Disadvantages:

- Although the study shows that the feature generation method outperforms the mean processing method, it is not clear how to interpret the results or the specific features that contribute to the accuracy of the models.
- If the data used in the study is biased or incomplete, the accuracy of the machine learning algorithms can be compromised, leading to unreliable results.

2.3 Multi-Branching Temporal Convolutional Network for Sepsis Prediction [3]

Sepsis is among the leading causes of morbidity and mortality in modern intensive care units. However, realworld medical data are often complexly structured with a high level of uncertainty (e.g., missing values, imbalanced data). In this paper, a novel predictive framework with Multi-Branching Temporal Convolutional Network (MB-TCN) to model the complexly structured medical data for robust prediction of sepsis is proposed.Experimental results show that MB-TCN outperforms existing methods that are commonly used in current practice.

Advantages:

- This framework effectively captures the temporal pattern and heterogeneous variable interactions.
- It also handles missing value and imbalanced data issues.

Disadvantages:

- This framework becomes complex when data becomes large.

2.4 Predicting Infections Using Computational Intelligence – A Systematic Review [4]

Infections encompass a set of medical conditions of very diverse kinds that can pose a significant risk to health, and even death. As with many other diseases, early diagnosis can help to provide patients with proper care to minimize the damage produced by the disease, or to isolate them to avoid the risk of spread. In this context, computational intelligence can be useful to predict the risk of infection in patients, raising early alarms that can aid medical teams to respond as quick as possible. In the study the most widely addressed infection is by far sepsis, followed by Clostridium difficile infection and surgical site infections. Most works use machine learning techniques, from which logistic regression, support vector machines, random forest and naive Bayes are the most common. SVM, logistic regression ,random forest and naïve bayes are used for early identification of infections which shows better accuracy. The implementation of computational intelligence in healthcare can be expensive, particularly if it involves the development of new technologies or the restructuring of existing healthcare systems.

Advantages:

- SVM, logistic regression ,random forest and naïve bayes are used for early identification of infections which shows better accuracy.

Disadvantages:

- The implementation of computational intelligence in healthcare can be expensive, particularly if it involves the development of new technologies or the restructuring of existing healthcare systems.

2.5 A Deep Learning-Based Sepsis Estimation Scheme [5]

The objective of this research is to design and implement a machine learning (ML) based technique that can predict cases of septic shock and extreme sepsis and assess its effects on medical practice and the patients. Data from the laboratory tests serve as the primary early indicator of septic shock by confirming the presence of toxins. The core deep learning network used is CNN architecture of Long short-term memory (LSTM). The values used for the alerting system were found to have no statistically significant difference in the context of different ICU wards. The patients should have had positive blood culture during their interaction with the hospital. Data from the laboratory tests serve as the primary early indicator of septic shock by confirming the presence of toxins. The core deep learning network used is CNN architecture of Long short-term memory (LSTM).

Advantages:

- The values used for the alerting system were found to have no statistically significant difference in the context of different ICU wards.

Disadvantages:

- The patients should have had positive blood culture during their interaction with the hospital.
- The layers become more and accuracy is less.

2.6 Prediction of sepsis patients using machine learning approach: A meta-analysis [6]

The study found that the machine learning prediction models performed better than the existing sepsis scoring systems such as SIRS, MEWS, SOFA, and qSOFA for identifying and predicting sepsis patients. Predicting sepsis patients using machine learning models could guide physicians to actively monitor and take preventive actions to improve the patients' condition. It would also identify patients most in need of medical support, reduce wasting healthcare resources, and increase the desired sensitivity or specificity, resulting in a decreased numbers of false alarms.

Advantages:

- The performance of the machine learning models was compared with other traditional scoring systems.
- Used different kinds of database including the MIMIC-III (v1.3) database.

Disadvantages:

- The study does not suggest which model is best for predicting sepsis patients.

2.7 Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III data base [7]

While the use of transthoracic echocardiography (TTE) in the ICU is rapidly expanding, the contribution of TTE to altering patient outcomes among ICU patients with sepsis has not been examined. This study was designed to examine the association of TTE with 28-day mortality specifcally in that population. In a general population of critically ill patients with sepsis, use of TTE is associated with an improvement in 28- day mortality. The MIMIC-III database was employed to identify patients with sepsis who had and had not received TTE. The statistical approaches utilized included multivariate regression, propensity score analysis, doubly robust estimation, the gradient boosted model, and an inverse probability-weighting model to ensure the robustness of our findings. This study was designed to examine the association of TTE with 28-day mortality specifcally in that population.

Advantages:

- The developed methodology can be a useful diagnostic tool for clinical decision support.
- The random forest-improved fruit fly optimization algorithm-kernel extreme learning machine was used to effectively diagnose the sepsis.

Disadvantages:

- Performing a TTE can be time-consuming, which can delay other necessary interventions in critically ill patients with sepsis.

2.8 A New Effective Machine Learning Framework for Sepsis Diagnosis [8]

The proposed method has got 81.6% recognition rate, 89.57% sensitivity, and 65.77% specificity. A new learning strategy was proposed to boost the performance of the kernel extreme learning machine, known as, chaotic fruit fly optimization, and two new mechanisms were introduced into the original a fruit fly optimization, including the chaotic population initialization and the chaotic local search strategy.They performed the feature selection using the random forest before the construction of the classification model. The final established model, random forest-improved fruit fly optimization algorithm-kernel extreme learning machine, was used to effectively diagnose the sepsis.

Advantages:

- This model is used effectively to diagnose the sepsis.
- More amount of data with huge parameter set is required.

Disadvantages:

- The proposed method has a lower specificity rate, which means that it is more likely to produce false positive results, leading to unnecessary treatment or interventions.

2.9 Learning representations for the early detection of sepsis with deep neural networks [9]

They aimed to construct early-stage sepsis detection models using deep learning algorithms and methodologies and to assess the viability and improvement of the novel deep learning approach against that of the regression method utilizing traditional "temporal feature extraction". They achieved enhanced performance with input to output layer that is fed forward in only direction neural networks utilizing extended short-term memory and enhanced performance with deep neural networks by comparing them to reference features.

Advatanges:

- The use of deep neural networks with input to output layer feeding in only one direction has made the detection process faster and more efficient.
- The use of deep learning algorithms and methodologies has shown enhanced performance in detecting early-stage sepsis compared to traditional temporal feature extraction.

Disadvantages:

- Due to the complexity of deep learning algorithms, reproducing the same results on different datasets or with different parameters can be challenging. This makes it difficult to evaluate and compare the performance of different models.

2.10 Predicting sepsis with a recurrent neural network using the MIMIC III database [10]

The adult patients who were admitted to the critical care unit but did not meet the criteria for sepsis at the time of admission (from the MIMIC III database) were examined. To evaluate the performance of the prediction, look at the sequence length provided to the machine learning algorithms at various times before the beginning of sepsis. Additionally, the effect of the definition of sepsis onset is looked into. In comparison to similar efforts in the field, they evaluated the model using a rather large and thus more representative patient sample. Additionally, the effect of the definition of sepsis onset is looked into. The research proved that, when comparing prediction performance, a recurrent neural network outperforms insight. But further research is required for better results.

Advantages:

- The study looked at the prediction performance of the model at various times before the onset of sepsis. This provides a better understanding of how early the prediction can be made and how much time is available for intervention.

Disadvantages:

- The study only evaluated the performance of the prediction based on a single definition of sepsis onset, which may not be applicable to all clinical settings

2.11 An ensemble machine learning model for the early detection of sepsis from clinical data [11]

In this paper, proposed an ensemble model (lightgbm, xgboost, and random forest) that incorporated boosting and bagging tree models were created to predict sepsis. On a subset of the inner test data, they compared the outcomes of the ensemble model and calculated each model's evaluation metrics. AUC of 0.792 and ACC of 0.727 were the best performance numbers attained offline. Finally, the suggested model was assessed using all available test sets. The usefulness score for LightGBM's single model was merely -0.036. In comparison to a single tree-based model, the ensemble model performed better since it made use of the pre-treatment data.

Advantages:

- By using boosting and bagging techniques, the proposed model can effectively handle and learn from the complexity of the data, which improves its generalizability and ability to predict new and unseen data.

Disadvantages:

- The performance of an ensemble model is highly dependent on the quality of the input data. Poor quality data can result in poor performance and inaccurate predictions.

2.12 Early Prediction of Sepsis for ICU Patients using Gradient Boosted Tree [12]

The proposed system in this paper intended to find, validate, and test potential machine-learning algorithms for the early prediction of sepsis. They performed on decision trees, random forests, ads, gradient-boosted trees, and multilayer perceptrons will be used to develop prediction models for the 15 hours before the onset of sepsis. But the models are less accurate.

Advantages:

- The study adds to the growing body of literature on sepsis prediction using machine-learning algorithms, contributing to the development of more accurate models in the future.

Disadvantages:

- The implementation of such algorithms in real-world clinical settings may face challenges in terms of acceptance, interpretation, and integration into clinical workflows.

Chapter 3

ANALYSIS AND DESIGN

This chapter includes the analysis of requirements for the proposed project. This chapter contains

- Functional Requirements.
- Non-Functional Requirements.

3.1 Functional Requirements

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 solve an issue [18]. Analyzing functional Requirements includes a number of processes. The Functional Requirements include:

Software Requirements
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. It's open-source software with a three-clause BSD licence. The word panel data is an econometrics term for data sets that comprise observations for the same persons over multiple time periods. Its moniker is a pun on the term "Python data analysis".

Seaborn:
Seaborn is an amazing visualization library 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. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables for better understanding of dataset.

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. 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 offers a viable open-source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications.

Google Colab:
Colab is a cloud-based notebook environment that is free to use. 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.

Hardware Requirements

- 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

3.2 Non-Functional Requirements

Non-functional requirements describe how a system must behave and establish constraints of its functionality [18]. This type of requirements is also known as the system's quality attributes. The Non-functional requirements of this project are:

Usability: Usability defines how difficult it will be for a user to learn and operate the system. It is assessed by using Efficiency of use, Intuitiveness.

Performance: Performance is a quality attribute that describes the responsiveness of the system to various user interactions with it. Poor performance leads to negative user experience. It also jeopardizes system safety when it is overloaded.

Availability: Availability is gauged by the period of time that the system's functionality and services are available for use with all operations. So, scheduled maintenance periods directly influence this parameter. And it's important to define how the impact of maintenance can be minimized. When writing the availability requirements, the team has to define the most critical components of the system that must be available at all time. You should also prepare user notifications in case the system or one of its parts becomes unavailable.

Scalability: Scalability requirements describe how the system must grow without negative influence on its performance. This means serving more users, processing more data, and doing more transactions. Scalability has both hardware and software implications. For instance, you can increase scalability by adding memory, servers, or disk space, can compress data, use optimizing algorithms, etc.

Chapter 4

SOFTWARE DESIGN

This chapter consists of the design of the software Life Cycle model diagrams and their detailed explanation. Design is about choosing the architecture and solutions appropriate to the problem

4.1 Software Development Lifecycle

Scrum Model

The project is divided into 4 sprints. The sprint backlogs are Feature Selection, Preprocessing, Classification using MLP Classifier, Classification using Adaboost,Gradient Boosting Classifier, Classfication using Guassian Naive Bayes, Classification using Linear Discriminant Analysis, Classification using Random Forest, Model Evaluation and developing a GUI. Priorities are assigned to each backlog and based on the priorities the backlogs are implemented. During the each backlog implementation, the daily scrum meetings are conducted that helps to know What work from last meet? What work done? Are there any roadblocks in your way and after the backlog complete, retrospection is done [16]. This process is repeated until all the backlogs are finished. The Figure 4.1.1 describes the lifecycle model used for the proposed model.

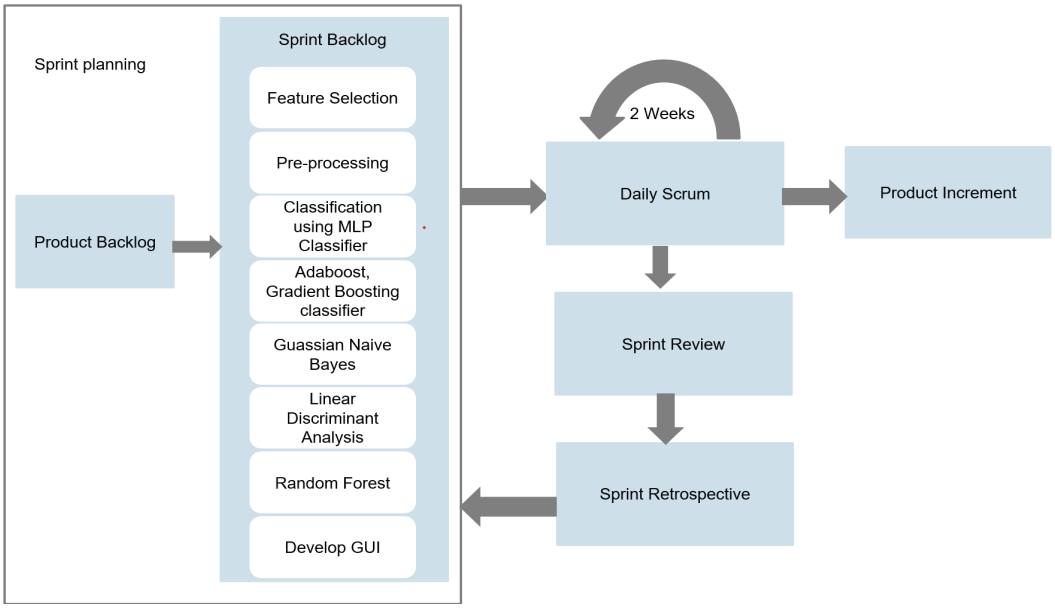


Figure 4.1.1: Scrum model

4.2 UML Diagrams

UML is a standard language for specifying, visualizing, constructing, and documenting the artefacts of software systems [17]. UML can be described as a general purpose visual modelling language to visualize, specify, construct and document software system. Although UML is generally used to model software systems, it is not limited within this boundary. It is also used to model non-software systems like process flow in a manufacturing unit etc. UML is not a programming language but tools can be used to generate code in various languages using UML diagrams. UML has a direct relation with object oriented analysis and design. The goal of UML can be defined as a simple modelling mechanism to model all possible practical systems in today's complex environment.

4.2.1 Activity Diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. From the website, the user sends the data of patients then it is classified by the trained model which got trained by the physioNet challenge dataset and best classifier of the considered classifiers. The response is displayed to user. The Activity diagram shows the overall flow of control in Figure 4.2.1.

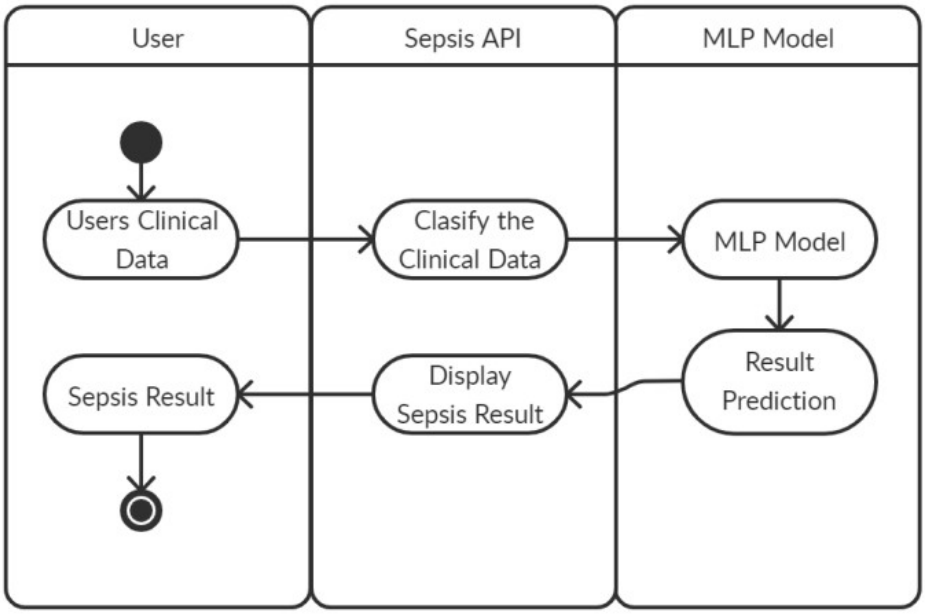


Figure 4.2.1: Activity Diagram

4.2.2 Use case Diagram

Use case diagrams are a way to capture the system's functionality and requirements in UML diagrams. It captures the dynamic behaviour of a live system. A use case diagram consists of a use case and an actor. From the website, the user sends the data of patients then it is classified by the trained model which got trained by the physioNet challenge dataset and best classifier of the considered classifiers. The response is displayed to user. The use case diagram for our project is shown in Figure 4.2.2.

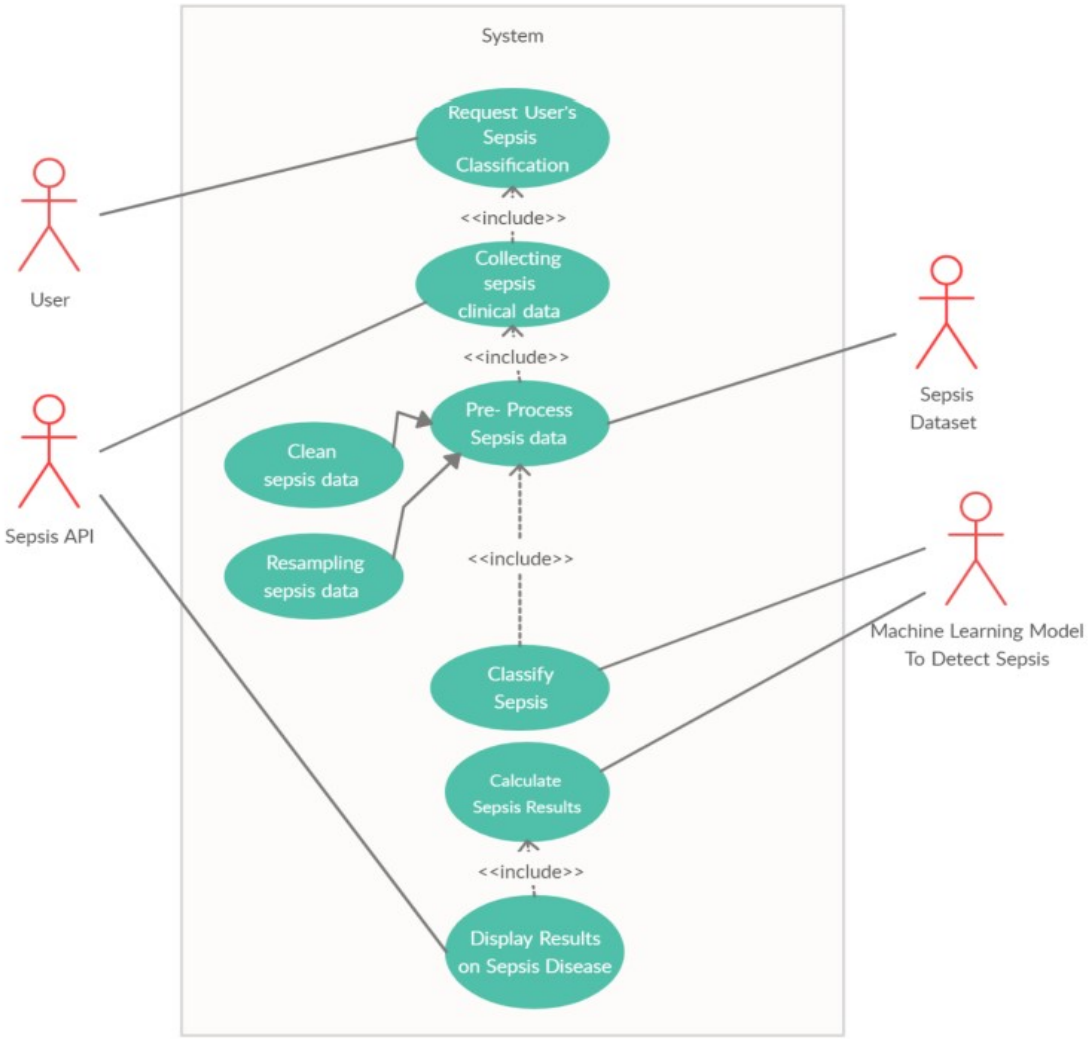


Figure 4.2.2: Use case Diagram

4.2.3 Sequence Diagram

A sequence diagram simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. From the website, the user sends the data of patients then it is classified by the trained model which got trained by the physioNet challenge dataset and best classifier of the considered classifiers. The response is displayed to user. For this project, the sequence diagram is shown in Figure. 4.2.3.

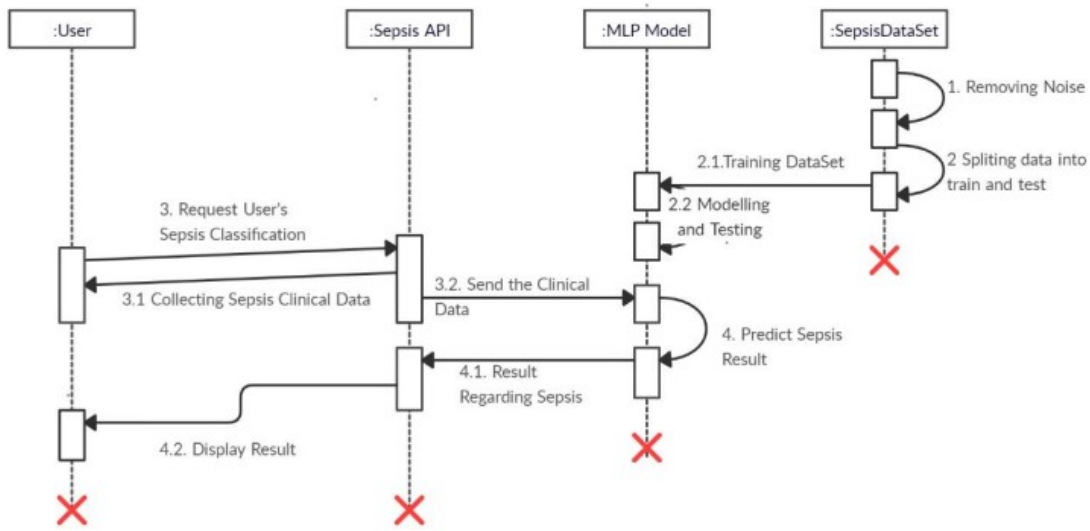


Figure 4.2.3: Sequence Diagram

Chapter 5

PROPOSED SYSTEM

This chapter includes the proposed system architecture along with the modules of methodology and dataset collection.

5.1 Architecture

The architecture for the proposed methodology is shown in Figure 5.1.1. It explains the process flow involved in the project.

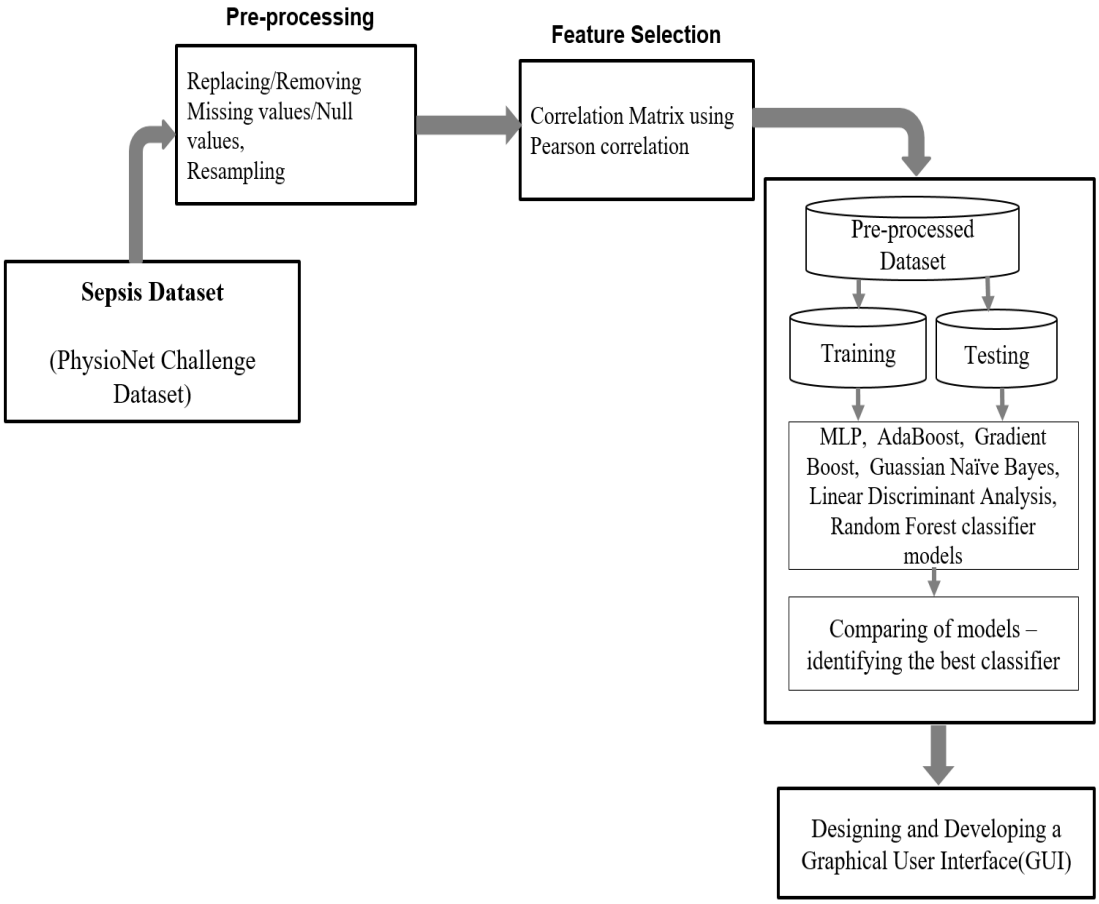


Figure 5.1.1: Architecture

5.2 Proposed Methodology

The primary intention of this project is to design and develop a technique for early detection of sepsis using different classifiers. The proposed technique involves four major steps, such as Pre-processing, Feature Selection, and Model Training using MLP, AdaBoost, Gradient boost, Guassian Naive Bayes, Linear Discriminant Analysis, Random Forest classifiers and Model Evaluation for identifying the best classifier based on accuracy, f1_score, AUC_ROC, Mean Absolute Error, Root Mean Squared Error and designing a website using Flask and integrating it with the best classifier. This project applies the proposed technique to PhysioNet challenge Dataset. Initially, the dataset is pre-processed i.e, replacing or removing the missing or null values, standardizing and undersampling the dataset is implemented. Then feature selection is done using correlation matrix using Pearson correlation to remove features having high correlation. Training the dataset using the MLP, AdaBoost, Gradient boost, Gaussian Naive Bayes, Linear Discriminant Analysis, Random Forest models and comparing the models based on performance measures and identifying the best classifier for the dataset. Designing a user-friendly website to predict sepsis.

5.3 Dataset

In this project, PhysioNet Challenge dataset is used to identify the best classifier to predict sepsis at early stages of occurrences.

Physionet Challenge Data

The dataset is collected from the Physionet challenge [15] where the data is gathered from patients in ICU from 3 separate hospitals. A total of 40,336 patients' clinical data from two definite hospitals were shared with the members while 22,761 patients' clinical data from three definite hospitals were segregated as obscure test sets. Each patient's clinical data contained likely 40 measurements of vital signs, laboratory, and demographics data. Each file has data separated with pipes in which each row represents a 1 hour's 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.

5.4 Algorithm

Step 1: Take the sepsis Dataset and install and import the required libraries.
Step 2: Pre-process the dataset by removing or replacing the null or missing values and re-sample the dataset so as to balance the dataset.
Step 3: Perform Feature Selection on the dataset considering correlation matrix using pearson correlation to remove the features that are having high correlation.
Step 4: Fit the data using fit(X,Y) where X, Y are input and output labels respectively.
Step 5: Take this preprocessed dataset and isolate it into Training and Testing dataset as X_train and Y_train and X_test, Y_test respectively,
Step 6: For each classifier in [MLP, Adaboost, Guassian Naive bayes, Gradient boosting, Linear discriminant Analysis, Random Forest] train each model with the Training dataset(X)
Step 7: Fit the train information into each classifier using fit(X,Y) where X and Y are input and output label respectively.
Step 8: Validate the prepared model with Test dataset.
Step 9: Identify the best classifier among the considered classifiers based on the performance measures accuracy, precison, recall,F1_Score AUC-ROC, Mean Absolute Error, Root Mean Square Error.
Step 10: Designing a GUI using Flask in which user can upload text files having values for the important features and predict whether the patient is having sepsis or not.

Chapter 6

IMPLEMENTATION

This chapter presents the coding part for all the modules involved in our project.

6.1 Pre-processing

For the raw dataset imputation is done to fill the missing values. In the further modified train dataset, removed the feature columns that are having null values greater than 25%. One hot encoding is implemented for gender column, Standard normalization for the remaining columns is implemented and NA values are dropped for the dataset. The number of rows with sepsis labels as 0 and 1 are not equal. So undersampling the dataset is done in order to balance the dataset. It is shown in Figure 6.1.1.

```
df_train_impute = df_train_mod.copy()
columns_impute = list(df_train_impute.columns)
grouped_by_patient = df_train_impute.groupby('Patient_ID'
,group_keys=False)
df_train_impute = grouped_by_patient.apply(lambda x: x.bfill
().ffill())
df_train_impute.head()
null_values = df_train_impute.isnull().mean()*100
null_values = null_values.sort_values(ascending=False)
null_values
null_col = ['TroponinI', 'Bilirubin_direct', 'AST',
'Bilirubin_total', 'Lactate', 'SaO2', 'FiO2','Unit',
'Patient_ID']
df_train_impute = df_train_impute.drop(columns=null_col)
df_train_impute.columns
one_hot = pd.get_dummies(df_train_impute['Gender'])
df_train_impute = df_train_impute.join(one_hot)
df_train_impute = df_train_impute.drop('Gender', axis=1)
df_train_impute.head()
columns_normalized = ['MAP', 'BUN', 'Creatinine', 'Glucose',
```

```
'WBC','Platelets']
for i in columns_normalized:
  df_train_impute[i] = np.log(df_train_impute[i]+1)
df_train_impute.head()
scaler = StandardScaler()
df_train_impute[['HR', 'O2Sat', 'Temp', 'MAP', 'Resp', 'BUN'
,'Chloride','Creatinine', 'Glucose', 'Hct', 'Hgb', 'WBC',
'Platelets' ]] = scaler.fit_transform(df_train_impute[['HR',
'O2Sat',  'Temp', 'MAP', 'Resp', 'BUN', 'Chloride',
 'Creatinine', 'Glucose', 'Hct', 'Hgb', 'WBC', 'Platelets'
 ]])
df_train_impute.head()
df_train_impute = df_train_impute.dropna()
null_values = df_train_impute.isnull().mean()*100
null_values
df_train_impute.head(5)
majority_class = df_train_impute[df_train_impute
['SepsisLabel'] == 0]
minority_class = df_train_impute[df_train_impute
['SepsisLabel'] == 1]
majority_class_subset = majority_class.sample
(n=2*len(minority_class),replace=True)
df_train_impute = pd.concat([majority_class_subset,
minority_class])
df_train_impute.head(5)
```

	Hour	HR	O2Sat	Temp	MAP	Resp	BUN	Chloride	Creatinine	Glucose	Hct	Hgb	WBC	Platelets	Age	HospAdmTime	ICULOS	SepsisLabel	0	1
696446	10	-0.232040	0.244815	0.464976	-1.191705	-1.439562	-0.898732	-0.957185	-0.410796	0.209533	-0.516838	-0.534709	-0.670346	0.771709	56.76	-0.01	11	0	1	0
222845	27	-0.114791	-0.065399	0.236741	0.464955	0.600192	0.256318	0.321441	0.341885	1.410288	-0.331683	-0.475794	-0.114851	0.728473	57.38	-17.34	28	0	1	0
328287	24	0.178331	-0.065399	1.163110	-0.180581	0.971056	-0.311977	-1.139846	-0.135134	0.234031	-0.249392	-0.357965	-0.180333	0.553254	58.22	-3.63	29	0	0	1
782653	8	-1.170030	-0.996040	-0.353987	-0.157228	-0.883265	-0.311977	-0.591863	-0.269336	-0.822764	0.429508	0.761415	-0.643282	0.699107	66.01	-0.03	10	0	1	0
620294	12	0.119706	0.244815	1.431622	0.358852	-0.141537	-0.521082	-0.591863	-0.410796	0.006646	-1.524902	-1.418430	0.262467	-0.103799	49.84	-32.22	13	0	0	1

Figure 6.1.1: Pre-processed dataset

6.2 Feature Selection

For the pre-processed dataset, generated the correlation matrix and removed the features that are having high correlation and drop those feature columns from the train dataset it is shown in Figure 6.2.1.

```
def corr_matrix(df):
    corr = df.corr()
    mask = np.triu(np.ones_like(corr, dtype=bool))
    f, ax = plt.subplots(figsize=(40,40))
    cmap = sns.diverging_palette(220, 10, as_cmap=True)
    sns.heatmap(corr, mask=mask, cmap="Paired", vmax=.3,
    center=0, square=True, linewidths=.5, cbar_kws
    ={"shrink": .5})
corr=df_train.corr()
plt.figure(figsize=(30,12))
sns.heatmap(corr,annot=True,cmap='coolwarm')
```

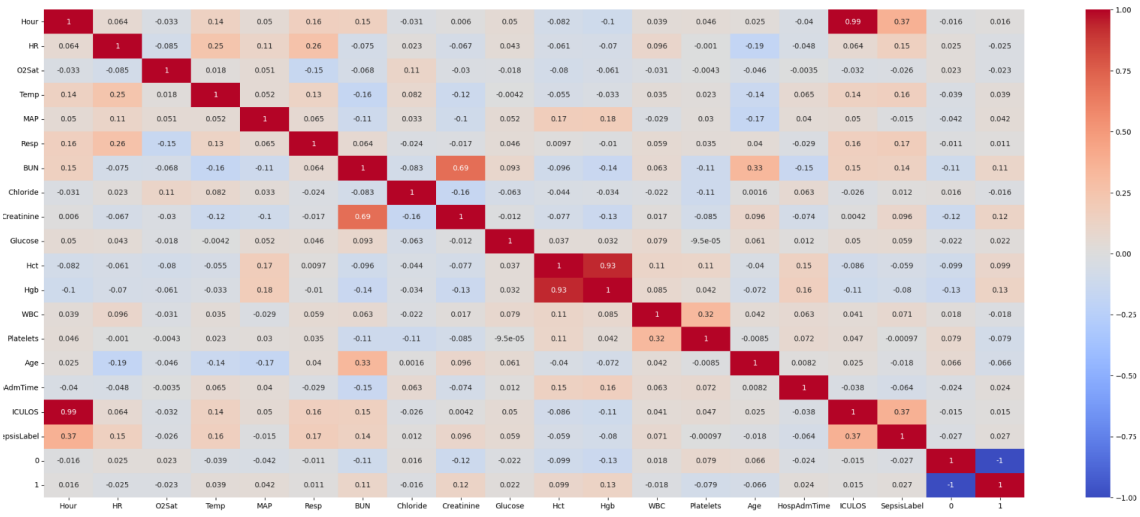


Figure 6.2.1: Correlation Matrix after removing high correlation features in dataset

6.3 Model Training

This balanced dataset is considered as the cleaned dataset and this dataset is divided into training and testing with test_size 0.2. Trained the cleaned dataset using MLP, Adaboost, Gradient Boost, Guassian Naive Bayes, Linear Discriminant Analysis, Random Forest Classifier.

```
classifiers = [
    MLPClassifier(
    activation='tanh',
    solver='lbfgs',
    early_stopping=False,
    hidden_layer_sizes=(40,10,10,10,10,2),
    random_state=1,
    batch_size='auto',
    max_iter=30,
    learning_rate_init=1e-5,
    tol=1e-4,
),
    AdaBoostClassifier(),
    GradientBoostingClassifier(),
    GaussianNB(),
    LinearDiscriminantAnalysis(),
    RandomForestClassifier(n_estimators=300, random_state=0)]
for clf in classifiers:
    clf.fit(X_train, y_train)
    name = clf.__class__.__name__
    print("="*30)
    print(name)
    print('****Results****')
    clf.fit(X_train, y_train)
    predictions = clf.predict(X_test)
    evaluate_model(y_test,predictions)
print("="*30)
```

6.4 Model Evaluation

Measuring the performance of the models based on their accuracy, F1_score, AUC-ROC, Mean Absolute Error, Root Mean Squared Error and identified the best classifiers among the trained models. The best classifier among the considered classifiers is Random Forest Classifier. It is shown in Figure 6.4.1.

```
def evaluate_model(y_true,y_pred):
  accuracy = accuracy_score(y_true, y_pred)
  print("Accuracy:", accuracy)
  precision = precision_score(y_true, y_pred)
```

```
print("Precision:", precision)
recall = recall_score(y_true, y_pred)
print("Recall:", recall)
f1 = f1_score(y_true, y_pred)
print("F1 Score:", f1)
auc = roc_auc_score(y_true, y_pred)
print("AUC-ROC:", auc)
mae = mean_absolute_error(y_true, y_pred)
print("Mean Absolute Error:", mae)
rmse = np.sqrt(mean_squared_error(y_true, y_pred))
print("Root Mean Squared Error:", rmse)
cm = confusion_matrix(y_true, y_pred)
sns.heatmap(cm, annot=True, fmt='d')
plt.show()
```

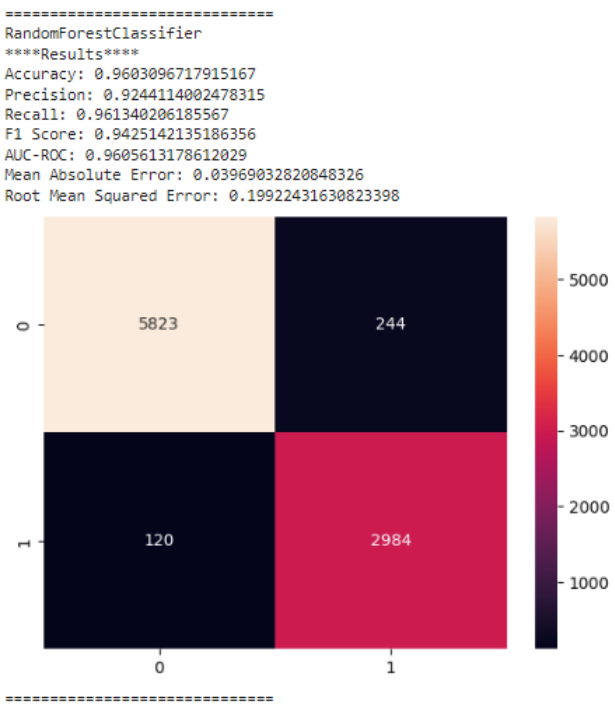


Figure 6.4.1: Performance Measures - Random Forest classifier

Developed a GUI using Flask to predict sepsis in which input files are considered as text files with 16 parameters of lab results of patients.

```
app = Flask(__name__)
model = joblib.load(open('model.pkl', 'rb'))

app.config['UPLOAD_EXTENSIONS'] = ['.txt']
app.config['UPLOAD_PATH'] = 'static/files/'
```

```
@app.route('/')
def home():
    print("home")
    return render_template('index.html')

@app.route('/predict',methods=['POST'])
def predict():
    print("predict")
    uploaded_file = request.files['file']
    print("file uploaded")
    filename = secure_filename(uploaded_file.filename)
    print(filename)
    if filename != '':
        file_ext = os.path.splitext(filename)[1]
        if file_ext not in app.config['UPLOAD_EXTENSIONS']:
            abort(400)
        uploaded_file.save(os.path.join(app.config
        ['UPLOAD_PATH'], filename))
    d = {}
    with open(os.path.join(app.config['UPLOAD_PATH'], filename))
    as f:
        for line in f:
            (key, val) = line.split()
            d[key] = val
    input = [float(i) for i in d.values()]
    prediction = model.predict([input])
    output = prediction[0]
    return render_template('results.html',predicted_image=
    'Prediction Result: {}'.format(output))
```

Chapter 7

TESTING

This chapter includes the testing of the proposed system. Unit testing is performed on each module.

1. **Test Case 1: Feature selection and Pre-processing**

Project Name: Sepsis Prediction using Random Forest Classifier						
Test case Id: 1			Test Designed By: Anupama			
Test Priority: High			Test Designed Date: 25-03-2023			
Module Name: Feature Selection & Pre-processing			Test Executed By: Lahari			
Test Title: Feature Selection & Pre-processing			Test Executed Date: 25-03-2023			
Description: Feature Selection & Preprocessing						
Pre-Conditions: Developer should import the libraries provide the raw data						
Stage	Test Steps	Test Data	Expected Result	Actual Result	Status (Pass/-Fail)	Remarks
1	Removing Null Values and Resampling	Raw data is loaded	Dataset is cleaned	Dataset is cleaned	Pass	Nil
2	Feature Selection	Pre-processed Dataset	16 important features are extracted in dataset	16 important features are extracted in dataset	Pass	Nil
Post-Conditions: Dataset is cleaned						

2. **Test Case 2: Predicting the sepis label**

Project Name: Sepsis Prediction using Random Forest Classifier.						
Test case Id: 2			Test Designed By: Anupama			
Test Priority: High			Test Designed Date: 25-03-2023			
Module Name: Predicting the sepsis label			Test Executed By: Lahari			
Test Title: Finding the label using RFC			Test Executed Date:25-03-2023			
Description: Finding the sepsis label						
Pre-Conditions: The cleaned dataset is used						
Stage	Test Steps	Test Data	Expected Result	Actual Result	Status (Pass/-Fail)	Remarks
1	Predicting the label	Cleaned dataset.	No Sepsis	No Sepsis	Pass	Nil
Post-Conditions: Getting the label of maximum probability						

Chapter 8

RESULTS

This chapter includes the results of the proposed system. The output screenshots of the project results are included in this chapter

8.1 Output Screenshots of models

The performance measures of MLP Classifer is shown in Figure 8.1.1. The accuracy is 75%.

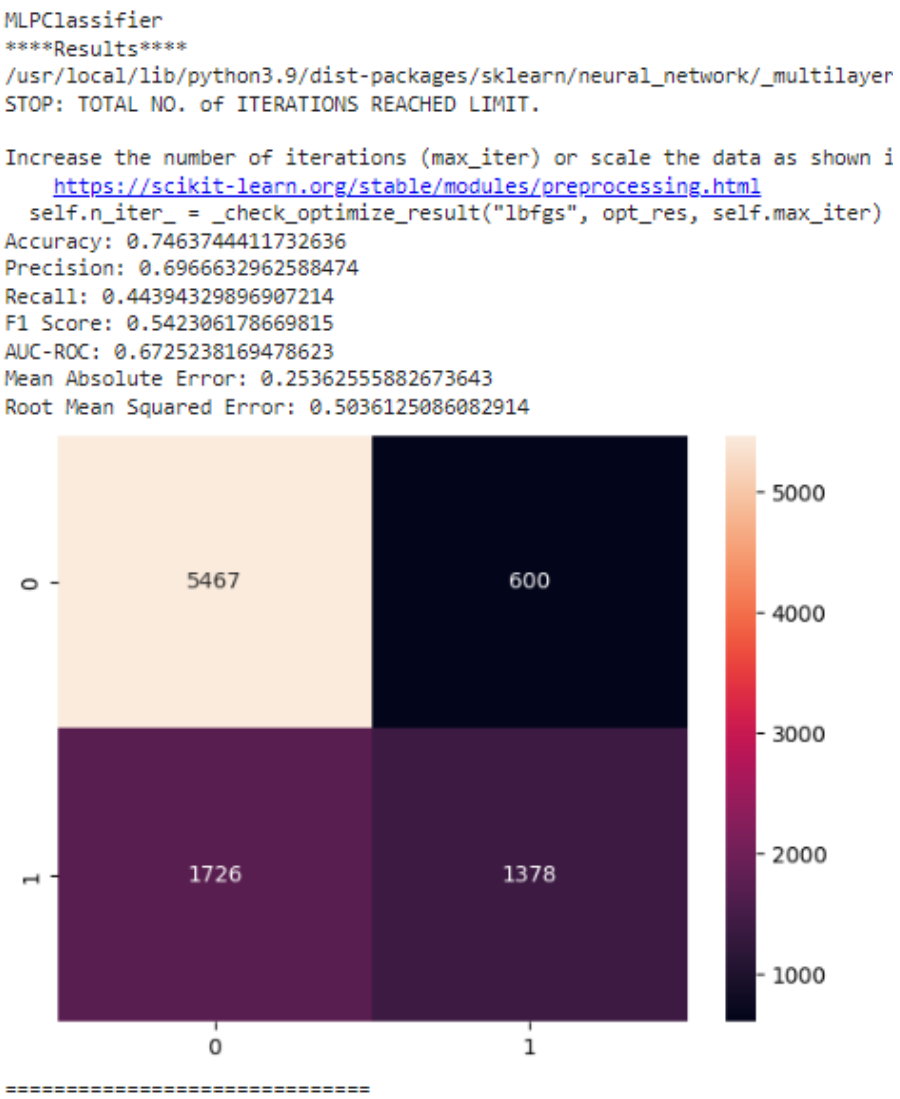


Figure 8.1.1: MLP Classifier's Performance

The performance measures of Adaboost Classifer is shown in Figure 8.1.2. The accuracy is 77%.

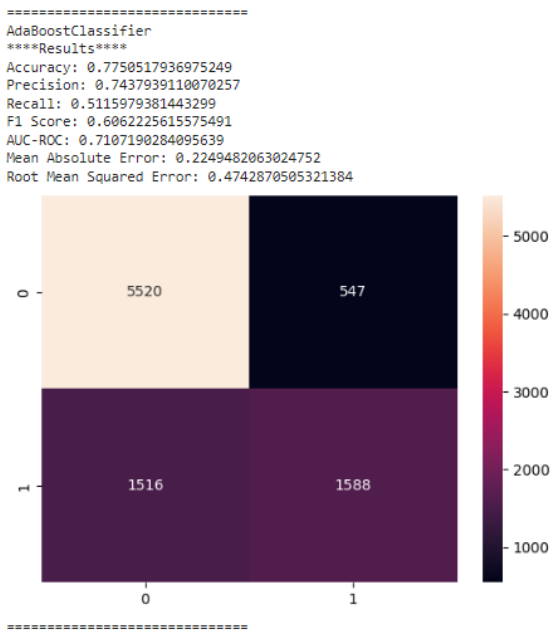


Figure 8.1.2: Adaboost Classifier's Performance

The performance measures of Gradient Boost Classifer is shown in Figure 8.1.3. The accuracy is 78%.

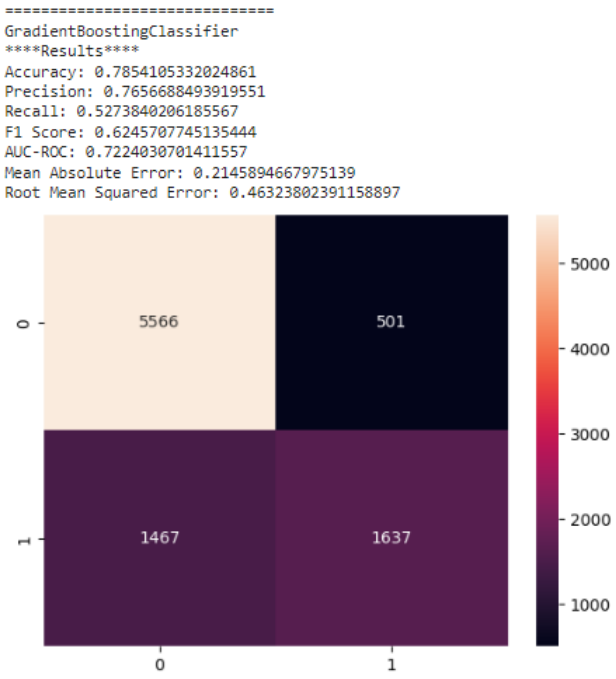


Figure 8.1.3: Gradient Boost Classifier's Performance

The performance measures of Guassian NB Classifer is shown in Figure 8.1.4. The accuracy is 74%.

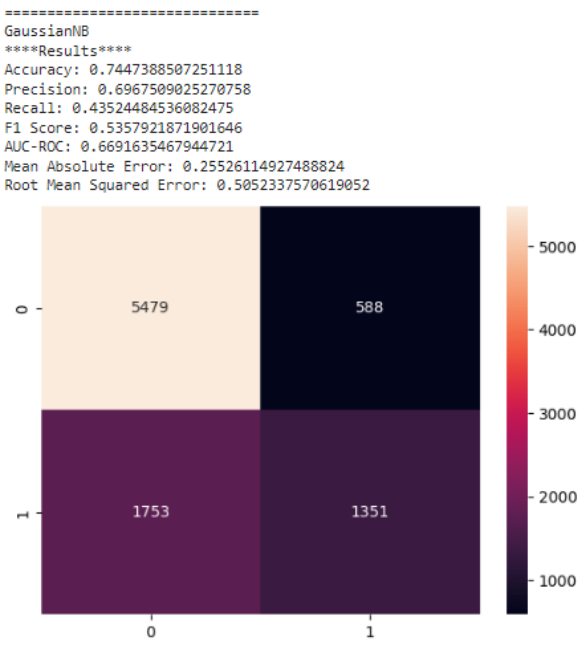


Figure 8.1.4: Guassian NB Classifier's performance

The performance measures of LDA is shown in Figure 8.1.5. The accuracy is 75%.

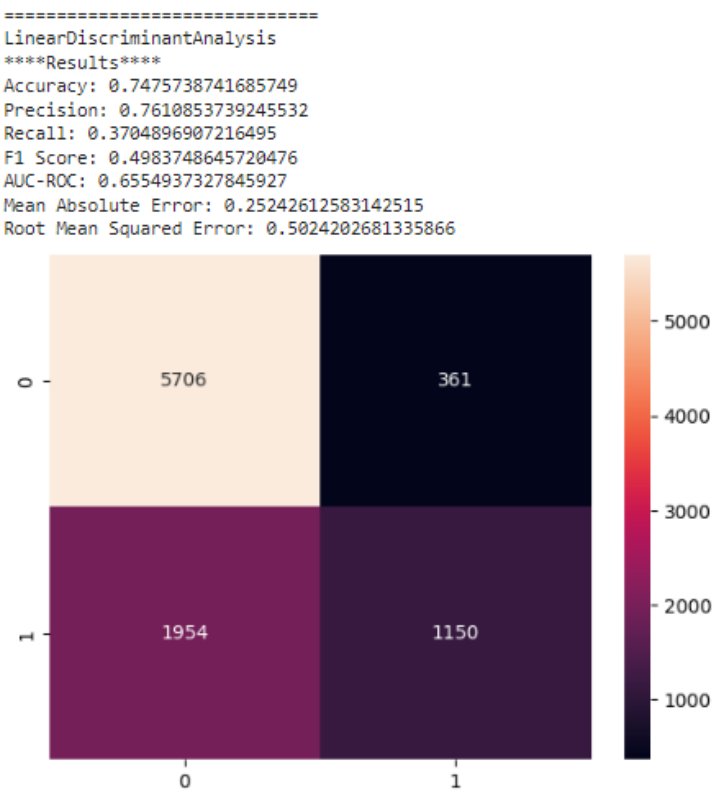


Figure 8.1.5: LDA Classifier's performance

The performance measures of Random Forest is shown in Figure 8.1.6. The accuracy is 96%.

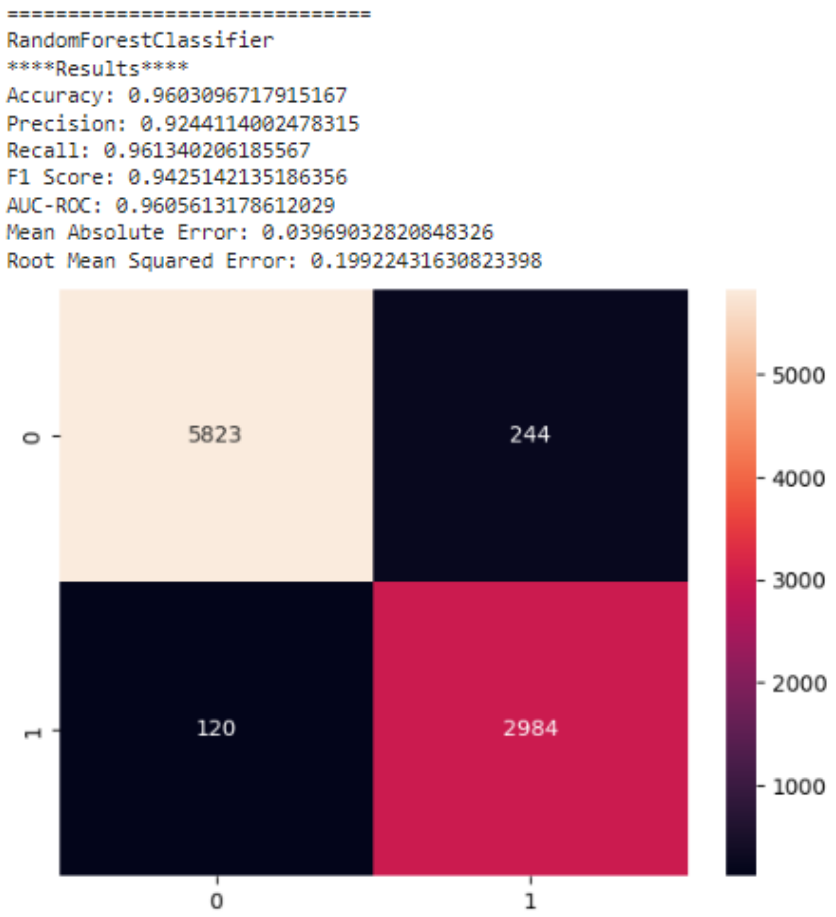


Figure 8.1.6: Random Forest classifier's performance

Among all the considered classifiers, Random Forest is the best classifier and it is considered as the best based on its performance measures. This classifier model has accuracy 96%. Table 8.1 shows the performance measures of all considered classifiers.

Classifier	Accuracy	F1_Score	AUC-ROC	MAE	RMSE
MLP Classifier	75%	54%	67%	25%	50%
Adaboost Classifier	77%	61%	71%	22%	47%
Gradient Boosting Classifier	78%	62%	72%	21%	46%
Guassian NB Classifier	74%	53%	67%	25%	50%
LDA Classifier	74%	50%	65%	25%	50%
Random Forest Classifier	96%	94%	96%	3%	19%

Table 8.1: Model Evaluation of classifiers

8.2 Output Screenshots of GUI

The user interface is shown in Figure 8.2.1. Here the input files accepted are text file formats which have the values of the 16 parameters of the lab results of the patient that are important in the dataset.

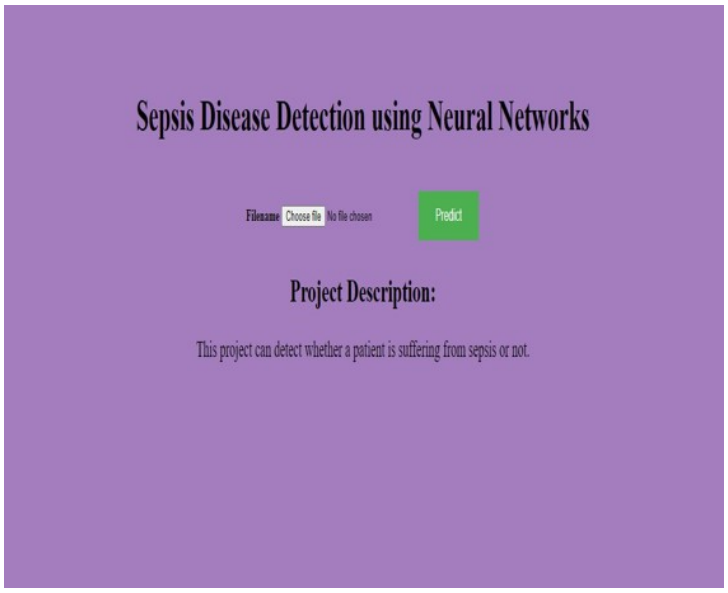


Figure 8.2.1: Developed GUI

If the patient is suffering from sepsis disease in GUI it is shown as in Figure 8.2.2 as the person is affected by sepsis.

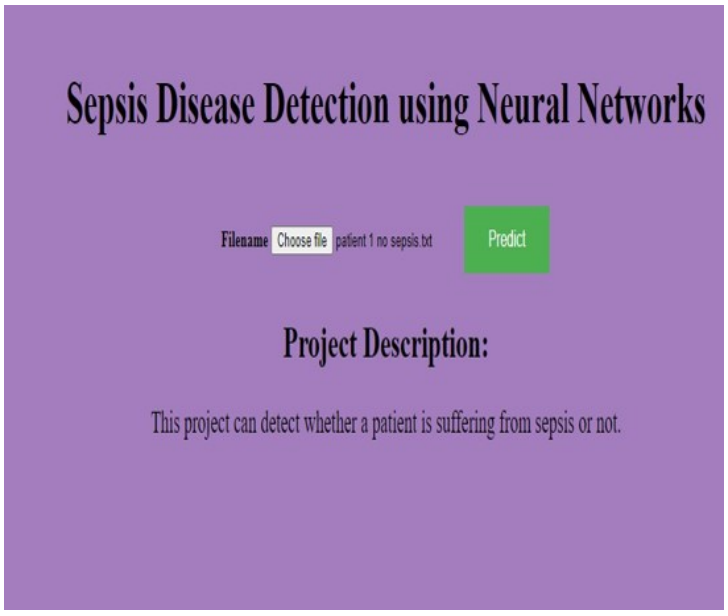


Figure 8.2.2: GUI result obtained when sepsis is present

If the patient is affected by sepsis disease in GUI it is shown as in Figure 8.2.3 as everything looks good.

Figure 8.2.3: GUI result obtained when no sepsis

Chapter 9

CONCLUSION AND FUTURE WORK

This chapter includes the conclusion of the project and future work.
Sepsis is a hazardous condition brought by an infection of the body. So as to prevent fungi, virus or bacteria, the body generally discharges the chemicals into the circulatory system. Sepsis happens as the body responds to these chemicals out of control, which induces changes that can affect the structures of many organs. This project has presented a description about Sepsis. The symptoms of the disease, signs, complications, and treatment for the disease are presented. This project also presents the detection of this disease at early stages with higher accuracy using different classifiers and found Random Forest classifier as the best among the considered classifiers based on performance measures which has accuracy 96%. In future, we would like to enhance the application by making it customer specific and deploying this model in a hospital website and help the doctors detect any early signs of the disease and also predicting which type of sespis and how much the patient is affected by the disease.

REFERENCES

[1] Liu, S., Fu, B., Wang, W., Liu, M., Sun, X. (2022). Dynamic sepsis prediction for intensive care unit patients using XGBoost-based model with novel time-dependent features. IEEE Journal of Biomedical and Health Informatics.

[2] Zhao, X., Shen, W., Wang, G. (2021). Early prediction of sepsis based on machine learning algorithm. Computational Intelligence and Neuroscience, 2021.

[3] Wang, Z., Yao, B. (2021). Multi-branching temporal convolutional network for sepsis prediction. IEEE Journal of Biomedical and Health Informatics, 26(2), 876-887.

[4] Baldominos, A., Puello, A., Oğul, H., Aşuroğlu, T., Colomo-Palacios, R. (2020). Predicting infections using computational intelligence–a systematic review. IEEE Access, 8, 31083-31102.

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

[6] Islam, M. M., Nasrin, T., Walther, B. A., Wu, C. C., Yang, H. C., Li, Y. C. (2019). Prediction of sepsis patients using machine learning approach: a meta-analysis. Computer methods and programs in biomedicine, 170, 1-9.

[7] 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.

[8] 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.

[9] 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.

[10] Scherpf, M., Gräßer, F., Malberg, H., Zaunseder, S. (2019). Predicting sepsis with a recurrent neural network using the MIMIC III database. Computers in biology and medicine, 113, 103395.

[11] Fu, M., Yuan, J., Lu, M., Hong, P., Zeng, M. (2019, September). An ensemble machine learning model for the early detection of sepsis from clinical data. In 2019 Computing in Cardiology (CinC) (pp. Page-1). IEEE.

[12] Ying, T. X., Abu-Samah, A. (2022, June). Early Prediction of Sepsis for ICU Patients using Gradient Boosted Tree. In 2022 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS) (pp. 78-83). IEEE.

[13] Available Online: https://becominghuman.ai/multi-layer-perceptron-mlp-models-on-real-world-banking-data-f6dd3d7e998f, Last Access: March 19, 2023.

[14] Available Online: https://www.geeksforgeeks.org/random-fores t-regression-in-python/, Last Access: March 21, 2023.

[15] Dataset Accessed From: https://physionet.org/content/challenge-2019/1.0 .0/, Last Access: March 19, 2023.

[16] Available Online: https://www.geeksforgeeks.org/scrum-software-develop ment/, Last Access: March 21, 2023.

[17] Available Online: https://www.geeksforgeeks.org/unified-modeling-langua ge-uml-introduction/, Last Access: March 21, 2023.

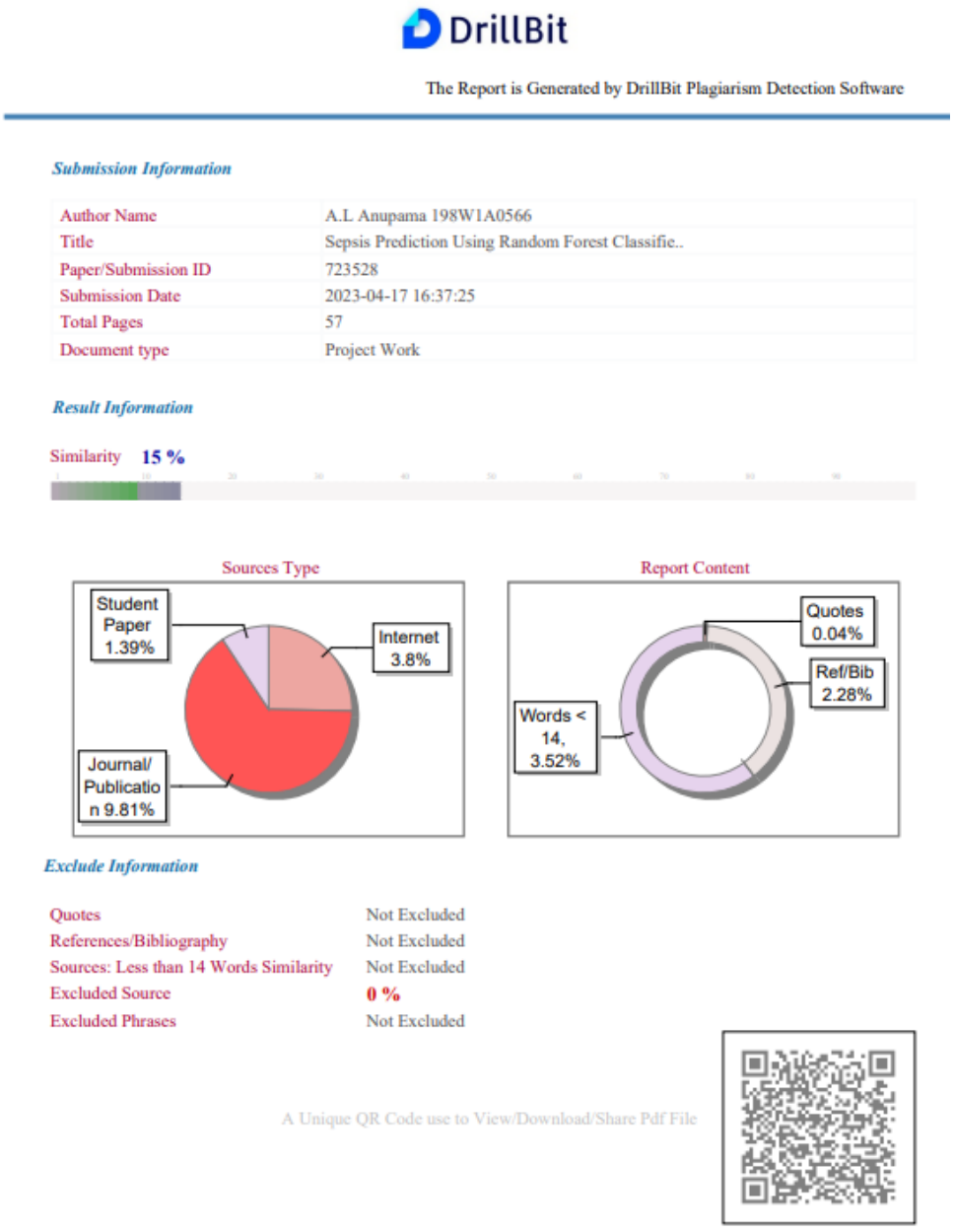
[18] Available Online: https://www.geeksforgeeks.org/functional-vs-non-funct ional-requirements/, Last Access: March 21, 2023.

PUBLICATION DETAILS

[1] Lalitha Anupama Annavarapu, A.V.L Lahari, S. Babu, "Prediction Of Sepsis using Machine Learning Algorithms", Submitted to International Journal of Online and Biomedical Engineering iJOE (Under Review).

APPENDIX - A
REPORT PLAGIARISM

The document entitled "Sepsis Prediction Using Random Forest Classifier" has plagiarism of 15% of similarity index.



DrillBit

The Report is Generated by DrillBit Plagiarism Detection Software

Submission Information

Author Name	A.L Anupama 198W1A0566
Title	Sepsis Prediction Using Random Forest Classifie..
Paper/Submission ID	723528
Submission Date	2023-04-17 16:37:25
Total Pages	57
Document type	Project Work

Result Information

Similarity 15 %

Sources Type

Report Content

Exclude Information

Quotes	Not Excluded
References/Bibliography	Not Excluded
Sources: Less than 14 Words Similarity	Not Excluded
Excluded Source	0 %
Excluded Phrases	Not Excluded

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

15	71	B
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	Jamia Milia Islamia University Thesis Published in inflibnet - www.inflibnet.ac.in	7	Publication
2	REPOSITORY - Submitted to Jawaharlal Nehru Technological University (H) on 2023-02-25 11-20	<1	Student Paper
3	A THIRD GENERATION DESIGN FOR THE AUTOMATED TELLER MACHINE OPERA By 16C91D5509 - 2019, JNTUH	<1	Student Paper
4	REPOSITORY - Submitted to Jawaharlal Nehru Technological University (H) on 2023-02-25 12-42	<1	Student Paper
5	www.frontiersin.org	<1	Internet Data
6	qdoc.tips	<1	Internet Data
7	REPOSITORY - Submitted to Jawaharlal Nehru Technological University (H) on 2023-02-24 18-34	<1	Student Paper
8	bg.copernicus.org	<1	Internet Data
9	www.ijariit.com	<1	Internet Data
10	www.frontiersin.org	<1	Internet Data
11	educationdocbox.com	<1	Internet Data
12	www.dx.doi.org	<1	Publication

13	members.cbio.mines-paristech.fr	<1	Internet Data
14	link.springer.com	<1	Internet Data
15	www.scribd.com	<1	Internet Data
16	dspace.nwu.ac.za	<1	Publication
17	www.hindawi.com	<1	Internet Data
18	Characterizing pattern preserving clustering by Hu-2008	<1	Publication
19	univagora.ro	<1	Publication
20	Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Develo by Keshavarz-2020	<1	Publication
21	The Promise and Perils of Wearable Physiological Sensors for Diabetes by Schwartz-2018	<1	Publication
22	docplayer.net	<1	Internet Data
23	members.cbio.mines-paristech.fr	<1	Internet Data
24	Association Between Arterial Calcifications and Nonlacunar and Lacunar Ischemic by va-2014	<1	Publication
25	Ensemble models from machine learning an example of wave runup and co by Beuzen-2019	<1	Publication
26	www.doaj.org	<1	Publication
27	www.ncbi.nlm.nih.gov	<1	Internet Data
28	doc.lagout.org	<1	Publication
29	FORMULATION AND INVITRO EVALUATION OF IMMEDIATE RELEASE TABLETS OF NEVIRAPINE BY 18DH1S0308 YR 2020, JNTUH	<1	Student Paper

30	researchspace.ukzn.ac.za	<1	Publication
31	Stability of embeddings for pseudoconcave surfaces and their boundaries by Charle-2000	<1	Publication
32	1library.co	<1	Internet Data
33	docplayer.net	<1	Internet Data
34	moam.info	<1	Internet Data
35	Constructing SSLM Insights from Struggles over Womens Rights in Nepal by Becker-2015	<1	Publication
36	ijcsmc.com	<1	Publication
37	Managing the tug-of-war between supply and demand in the service industries by Gabrie-1997	<1	Publication
38	store.nolo.com	<1	Internet Data
39	Use of Fatty Acid Methyl Ester Profiles for Discrimination of Bacillus Cereus T-Strain Spores Grown on Different Medi	<1	Internet Data
40	www.uni-obuda.hu	<1	Publication
41	Assessment of supervised machine learning for atmospheric retrieval of exoplanet by Nixon-2020	<1	Publication
42	qdoc.tips	<1	Internet Data
43	The relationship between delay discounting and Internet addiction A systematic by Cheng-2021	<1	Publication
44	En las Manos de Dios in Gods Hands Religious and Other Forms of Coping Among by Abrado-Lanza-2004	<1	Publication
45	IEEE 2018 International Conference on Sustainable Information Engin, by Lestarini, Dinda R- 2018	<1	Publication

46	moam.info	<1	Internet Data
47	moam.info	<1	Internet Data
48	plosjournal.deepdyve.com	<1	Internet Data
49	www.dx.doi.org	<1	Publication
50	www.ijitee.org	<1	Publication
51	63-service.ru	<1	Internet Data
52	Bharathidasan University Thesis published in - www.inflibnet.com	<1	Publication
53	cran.r-project.org	<1	Internet Data
54	dochero.tips	<1	Internet Data
55	docplayer.net	<1	Internet Data
56	docplayer.net	<1	Internet Data
57	docplayer.net	<1	Internet Data
58	erj.ersjournals.com	<1	Internet Data
59	erj.ersjournals.com	<1	Internet Data
60	healthmanagement.org	<1	Publication
61	ijircce.com	<1	Publication
62	Imaging and CSF Studies in the Preclinical Diagnosis of Alzheimers Di by M-2007	<1	Publication
63	mdpi.comjournalacoustics	<1	Internet Data
64	members.cbio.mines-paristech.fr	<1	Internet Data

65	ojs.unm.ac.id	<1	Publication
66	opac.fkik.uin-alauddin.ac.id	<1	Publication
67	repositorioslatinoamericanos	<1	Publication
68	The impact of management control systems on organisational change and performanc by Nuhu-2019	<1	Publication
69	uir.unisa.ac.za	<1	Publication
70	www.sciencedirect.com	<1	Internet Data
71	www.thefreelibrary.com	<1	Internet Data

APPENDIX - B
CODE PLAGIARISM

The major project code has plagiarism of 19% of similarity index.

DrillBit

The Report is Generated by DrillBit Plagiarism Detection Software

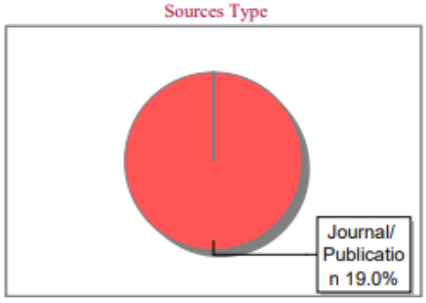
Submission Information

Author Name	A.L Anupama 198W1A0566
Title	Sepsis Prediction Using Random Forest Classifie..
Paper/Submission ID	723784
Submission Date	2023-04-18 11:42:09
Total Pages	4
Document type	Project Work

Result Information

Similarity **19 %**

Sources Type



Report Content

Exclude Information

Quotes	Not Excluded
References/Bibliography	Not Excluded
Sources: Less than 14 Words Similarity	Not Excluded
Excluded Source	**0 %**
Excluded Phrases	Not Excluded

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

19	2	B	A-Satisfactory (0-10%) B-Upgrade (11-40%) C-Poor (41-60%) D-Unacceptable (61-100%)
SIMILARITY %	MATCHED SOURCES	GRADE	

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1	dokumen.pub	16	Publication
2	dokumen.pub	3	Publication