DETECTION OF SEPSIS ON CLINICAL DATA USING MULTI - LAYER PERCEPTRON

A Mini Project report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Miss. Karadi Prathyusha	**(17071A05J9)**
Miss. Yelamarthi Meghana	**(17071A05N6)**
Miss. Yendluri Hari Chandana	**(17071A05O0)**
Miss. Yedla Shreshta Reddy	**(18075A0548)**

Under the guidance of

Mrs N. Venkata Sailaja

(Assistant Professor, VNR VJIET)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VNR VIGNANA JYOTHI INSTITUTE OF ENGINEERING & TECHNOLOGY

(An Autonomous Institute, NAAC Accredited With 'A++' Grade, NBA Accredited, Approved by AICTE, New Delhi, Affiliated to JNTUH)

VALLURUPALLI NAGESHWARA RAO VIGNANA JYOTHI
INSTITUTE OF ENGINEERING & TECHNOLOGY

(An Autonomous Institute)

Hyderabad – 500090

CERTIFICATE

This is to certify that **Miss. Karadi Prathyusha (17071A05J9), Miss. Yelamarthi Meghana (17071A05N6), Miss. Yendluri Hari Chandana (17071A05O0) and Miss. Yedla Shreshta Reddy (18075A0548)** have successfully completed their mini project work at Department of CSE, VNR VJIET, Hyderabad entitled **"DETECTION OF SEPSIS ON CLINICAL DATA USING MULTI-LAYER PERCEPTRON"** in partial fulfilment of the requirements for the award of B. Tech degree during the academic year 2020-21.

Mrs N. Venkata Sailaja
Assistant Professor & Internal Guide
Department of Computer Science
VNR VJIET

Mrs B.V. Kiranmayee
Associate Professor & HOD
Department of Computer Science
VNR VJIET

DECLARATION

We hereby declare that the mini project entitled **"DETECTION OF SEPSIS ON CLINICAL DATA USING MULTI -LAYER PERCEPTRON"** submitted in partial fulfilment of the requirements for award of the degree of Bachelor of Technology in Computer Science and Engineering at **VNR Vignana Jyothi Institute of Engineering and Technology,** affiliated to Jawaharlal Nehru Technological University, Hyderabad, is a bonafide report of the work carried out by us under the guidance and supervision of Mrs N. Venkata Sailaja (Assistant Professor), Department of CSE, VNRVJIET. To the best of our knowledge, this report has not been submitted in any form to any University/Institute for award of any degree or diploma.

Karadi Prathyusha	**Yelamarthi Meghana**	**Yendluri Hari Chandana**	**Yedla Sai Sreshta**
(17071A05J9)	**(**17071A05N6)	(17071A05O0)	(18075A0548)
IV B. Tech CSE	IV B. Tech CSE	IV B. Tech CSE	IV B. Tech CSE

ACKNOWLEDGEMENT

Behind every achievement lies an unfathomable sea of gratitude to those who activated it, without which it would ever never have come into existence. To them we lay the words of gratitude imprinting within us.

We are indebted to our venerable principal **Dr. C. D. Naidu** for this unflinching devotion, which lead us to complete this project. The support, encouragement given by him and his motivation lead us to complete this project. We are very much thankful to our H.O.D., **Mrs B.V. Kiranmayee** madam for extending her cooperation in doing this project.

We extend our heartfelt thanks to our guide, **Mrs N. Venkata Sailaja** ma'am, for her enthusiastic guidance throughout the course of our project. We also express our sincere thanks to our Mini project co-ordinators **Dr.A. Brahmananda Reddy and Mrs N. Venkata Sailaja** who extended their valuable support in helping us complete the project in a correct way.

Last but not the least, our appreciable obligation also goes to all staff members of Computer Science & Engineering Department and to our fellow classmates who directly or indirectly helped us.

Miss. Karadi Prathyusha (17071A05J9)
Miss. Yelamarthi Meghana (17071A05N6)
Miss. Yendluri Hari Chandana (17071A05O0)
Miss. Yedla Shreshta Reddy (18075A0548)

ABSTRACT

Sepsis is activated by the immune system present in your 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. For the patient the practicality of detecting sepsis disease occurrence in development is an important factor in the result. The primary goal of this work is to build, train and test a Multi-Layer Perceptron (MLP) model using data that is available in the form of electronic clinical health data and predicts outcome of class labels as sepsis or no-sepsis for unseen health records. The secondary goal is to compare the accuracy of the MLP against many other models like Ada-Boost, Gradient Boosting, etc. using the metrics accuracy and log loss.

INDEX

Contents Page. No

CHAPTER 6

CHAPTER 7

LIST OF FIGURES

1. INTRODUCTION

1.1 Introduction

1.1.1 Sepsis

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. 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.

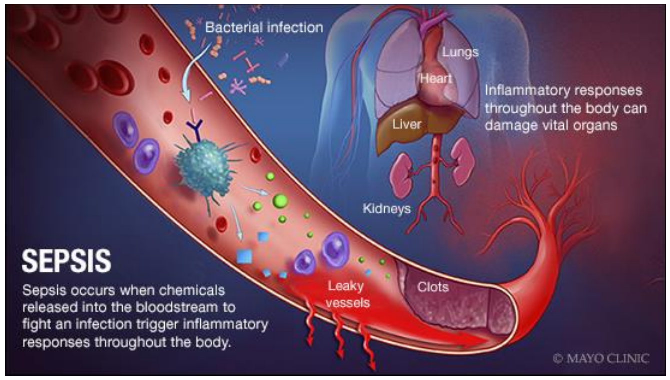


Figure 1.1.1 Image of sepsis

1.1.2 Sepsis Symptoms

Symptoms include:

Fever: temperature above 38ºC or even the temperature below normal i.e., 36ºC.

Heart rate: greater than 90 beats/min.

Breathing: higher than 20 breaths/min.

1.1.3 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.

1.1.4 Treatments for Sepsis:

Antibiotics via IV in order to fight infection, as the patients suffer from lower blood pressure with help of vasoactive medications in order to increase blood pressure. Insulin is given in order to control the sugar levels in blood. Corticosteroids are used to reduce inflammation caused due to bacteria in the blood. Painkillers are also given in order to bear pains due to treatment. At extreme stages the organs get infected and damaged, in such a case if the kidneys are affected then Dialysis treatment is given in order to sustain.

1.1.5 Statistical Information

"Population polls found low population perception of sepsis and its impact ranging from 14% in Brazil to 40% in Australia. Sadly, public awareness data is not accessible for India, "said Dilip Mathai, Dean of Apollo Medical Sciences & Research Institute (AIMSR), Hyderabad

A 2016 study reported that almost 30% of patients have been admitted to ICUs in India who came out to have sepsis; it also showed that one out of three of those patients died. Research has identified a heavy sepsis burden in pregnant mothers and the neonate. A study conducted in 2017 by leading journal Lancet recorded that communicable diseases (infections) led to a significant proportion of deaths in India.

1.2 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.

1.2.1 Drawbacks of Existing Systems

The drawbacks of the current system include the following:

- Gradient vanishing and exploding problems. Training an LSTM is a very difficult task. 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.

1.3 Proposed System

The primary intention of this research is to design and develop a technique for early detection of sepsis using Multi-Layer Perceptron. The proposed technique involves three major steps, such as pre-processing, feature importance, and classification. Initially, the data will be pre-processed using the resampling technique. The feature importance is done using Xgboost Algorithm. The proposed classifier, named Multi-Layer Perceptron Classifier gives the detection results on the sepsis. The technique we proposed is shown as the architecture - the block view of the modules of the proposed system (In figure 2). The implementation of the proposed technique will be in PYTHON. The system is evaluated in terms of Accuracy and Log Loss in order to show the performance of the technique. This measured performance of the technique which is proposed in this paper will be compared with that of the existing work.

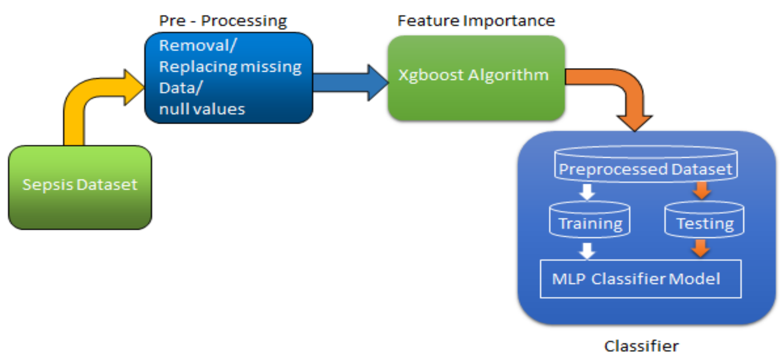


Fig 1.3 Proposed Technique - Block diagram

1.3.1 Advantages of Proposed System

The advantages of the Proposed System include the following:

- 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.

2. FEASIBILITY STUDY

A feasibility study involves taking a judgment call on whether a project is doable. The two criteria to judge feasibility are **cost required** and **value to be** delivered. A well-designed study should offer a historical background of the business or project, a description of the product or service, accounting statements, details of operations and management, marketing research and policies, financial data, legal requirements and tax obligations. Generally, such studies precede technical development and project implementation

.

A feasibility study evaluates the project's potential for success; therefore, perceived objectivity is an important factor in the credibility of the study for potential investors and lending institutions.

2.1 Technical Feasibility

Technical feasibility involves evaluation of the hardware and the software requirements of the proposed system.

In this project, the technology involved is Machine Learning. The language that is used to implement the concepts of Machine Learning is Python Programming and the tool that is used to execute the Python code is Jupyter Notebook (IPython notebook).

2.2 Economic Feasibility

Economic Feasibility helps in assessing the viability, cost, and benefits associated with projects before financial resources are allocated. This assessment typically involves a cost/ benefits analysis of the project.

The application is so designed that it requires minimal cost and eliminates costs as there would minimal need for manual work. The technologies used, help in understanding the user without any investment. As the machine will be trained it reduces the cost that is required to deploy the man power and also eliminates the problem of time consumption.

2.3 Legal Feasibility

The proposed system doesn’t conflict with legal requirements like data protection acts or social media laws. It ensures legal data access and gives prominence to data security.

2.4 Operational Feasibility

The application involves design-dependent parameters such as reliability, maintainability, supportability, usability, disposability, sustainability, affordability, and others.

It minimizes the drawbacks of the current system by building an application that automatically resolves the user queries and helps to analyse the user data.

2.5 Scheduling Feasibility

The project development took place in a timely process by understanding time schedules of the project and maintaining a good time line for project development.

3. LITERATURE SURVEY

Sepsis is an infection- chronic inflammatory disorder, typically results from the spread of a localized nidus to the systemic circulation, and both with very high deaths and morbidity levels correlated.

It is one of the most critical factors of in-hospital deaths. However, a credible way to predict septic origins remains elusive. Early and reliable projections of sepsis will enable further agitation and tailored treatment while antimicrobial stewardship is preserved. Established detection approaches are badly implemented and need laboratory test tests, sometimes overtime.

Recently, automated testing has proved to save lives. Tackling and researching enormous physiological observations continuously seen in ICU patients should immediately boost early position estimates, monitoring and essential disease treatment.

Under these conditions, beneficial local inflammatory processes, intervened by specific white platelets, for example, neutrophils and monocytes and the components they produce, and which are ordinarily present to control the spread of the irresistible centre, may grow their circles of movement into perilous fundamental irritation. The course by which a patient advances either to death or medical clinic release is notable and has been portrayed as a continuum from a state named foundational incendiary reaction disorder (SIRS) to progressive conditions of sepsis, serious sepsis, septic stun, numerous end-organ disappointment (MODS) and demise

Better health outcomes have been related to MLA. That's the primary randomized random experimental managed frame in septic reconnaissance to demonstrate observable comparisons and clinical deaths

On using The Dynamic Bayesian Networks, a time- probabilistic method of predicting a network utilizing patient data admitted into an emergency room, assessed the precision of diagnosis of sepsis in the first six hours after entry. The area under the curve was 0.915.

In the light of initial studies in comparison to HC and septic sequence sets, there were 42-gender markers that spoke of important intrinsic and diversified resistance capacities, cell cycling, differentiation of wireless connectivity, additional cell remodelling and immune modulation pathways. A LogitBoost algorithm has been used to construct a symptomatic learning guideline for predicting septic series outcomes. The accuracy was around 86%.

All factors which are significant for prediction (i.e., predictor variables) utilized for the examination show the distinction between both sepsis and non-sepsis patients. In patients, without sepsis, the mean temperature for hypothermia was considerably less. The risk of sepsis was 2,126 for patients with 38 ° C or higher temperatures.

A Model, which was built on combining both boosting and bagging tree models (lightgbm, xgboost, and random forest) have been built to predict on the basis of patient hourly data reports the best performance achieved was 79.2%.

In this analysis of proof of definition, AI proposes a close-by huge solution to overcome current CDRs as well as standard strategies for estimating ED patients with septic disease mortality in the hospital. The viability of this methodology should be tentatively evaluated and whether further research should turn this into better clinical results for high-risk patients with sepsis. The approaches created to support, for example, another model for detailed crash tests that can be robotized and applied for certain clinical results of a plot and submitted to EHRs for local clinical predictions.

Three specific approaches to describe the obligations of this document: (a) improved execution by utilizing field detail extraction, (b) check the capabilities to extract deep neural systems by correlating with reference highlights, and (c) Enhanced execution with LSTM, a neural network architecture feed for neural networks that are able to know patterns.

When using LSTM to diagnose the septic shock early, patients are identified by identical highlights and aim meanings up to 20 hours earlier than the Cox relative hazard model, with equal affectability and explicitly. This result is significant as early detection and treatment of the septic condition is necessary to improve the stamina capacity of the patient.

The LSSVM proposed was tested using a 5-fold cross-validation technique to execute with 2 separate kernels: the cubic-polynomial and the Gaussian radial base (RBF). The analysis revealed that LSSVM with RBF kert was a successful classifier to classify the development of sepsis syndrome with an accuracy of classification of 93.32 percent.

A thorough early learning calculation on multi-focus Danish information that focuses on time precision. The findings range from AUROC 0.856 (3 hours prior to the onset of sepsis) to AUROC 0.756 (24 hours prior to sepsis).

A meta-analysis of quantitative research is conducted to test the display of the septic learning pattern. This produced a pool of 0.89 (95%CI: 0.86-0.92); a responsiveness of 0.81 (95%CI: 0.80-0.71); a speciality of 0.72 (95%CI: 0.72-0.72) in the area that acknowledged the functioning bend (SAUROC). In this paper, they have compared most of the machine learning algorithms and declared that CNN-LSTM neural networks performed the best.

3.1 About Machine Learning

Machine learning is a buzzword these days, the reason for this is the huge amount of data production by applications and the increase of computation power. The term machine learning was first introduced by Arthur Samuel in 1959. We can define it in a summarized was as:

“Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed.”

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without beginning explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn from themselves.

3.2 Working of a Machine Learning Model

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of the predicted output depends up on the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it we just need to feed the data to generic algorithms, and with the help of these algorithms, the machine builds the logic as per the data and predicts the output. Machine learning has changed our way of thinking about the problem, Block diagram of machine learning algorithm is as follows

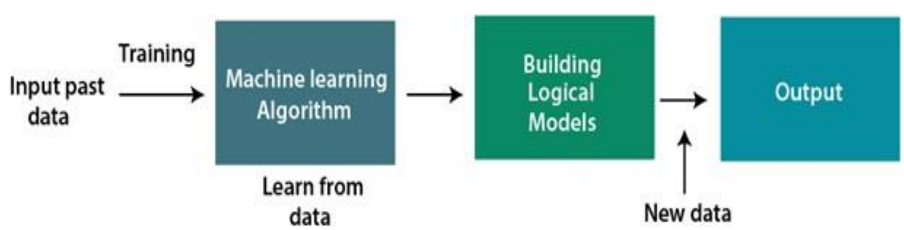


Figure 3.2 Block Diagram of Machine Learning Algorithm

3.3 Types of Machine Learning

Machine Learning is broadly classified into three types, they are:

- Supervised Learning
- Unsupervised Learning

Supervised Learning: The system creates a model using labelled data to understand the datasets and learn about each data, once the training and processing are done then we test the model by providing a sample data to check whether it is predicting the exact output or not. The goal of supervised learning is to map input data with the output data. The supervised learning is based on supervision, and it is the same as when a student learns things in the supervision of the teacher. The example of supervised learning is spam filtering. Supervised learning can be grouped further in two categories of algorithms. They are:

- Classification
- Regression

Unsupervised Learning: Unsupervised learning is a learning method in which a machine learns without any supervision. The training is provided to the machine with the set of data that has not been labelled, classified, or categorized, and the algorithm needs to act on that data without any supervision. The goal of unsupervised learning is to restructure the input data into new features or a group of objects with similar patterns. In unsupervised learning, we don't have a predetermined result. The machine tries to find useful insights from the huge amount of data. It can be further classified into two categories of algorithms. They are:

- Clustering

- Association

3.4 Applications of Machine Learning

Machine Learning is a diversified field and its applications are as follows:

- **Image Recognition:** Image recognition is one of the most common applications of machine learning. It is used to identify objects, persons, places, digital images, etc. The popular use case of image recognition and face detection is, Automatic friend tagging suggestion. Facebook provides us a feature of auto friend tagging suggestion. Whenever we upload a photo with our Facebook friends, then we automatically get a tagging suggestion with name, and the technology behind this is machine learning's face detection and recognition algorithm.

- **Speech Recognition:** While using Google, we get an option of "Search by voice," it comes under speech recognition, and it's a popular application of machine learning. Speech recognition is a process of converting voice instructions into text, and it is also known as "Speech to text", or "Computer speech recognition." At present, machine learning algorithms are widely used by various applications of speech recognition. Google assistant, Siri, Cortana, and Alexa are using speech recognition technology to follow the voice instructions.

- **Traffic Prediction:** If we want to visit a new place, we take help of Google Maps, which shows us the correct path with the shortest route and predicts the traffic conditions. It predicts the traffic conditions such as whether traffic is cleared, slow-moving, or heavily congested with the help of two ways: They are

- Real Time location of the vehicle form Google Map app and sensors

- Average time has taken on past days at the same time.

- **Product Recommendation:** Machine learning is widely used by various e-commerce and entertainment companies such as Amazon, Netflix, etc., for product recommendation to the user. Whenever we search for some product on Amazon, then we started getting an advertisement for the same product while internet surfing on the same browser and this is because of machine learning. Google understands the user interest using various machine learning algorithms and suggests the product as per customer interest. As similar, when we use Netflix, we find some recommendations for entertainment series, movies, etc., and this is also done with the help of machine learning.

- **Self-Driving Cars:** One of the most exciting applications of machine learning is self-driving cars. Machine learning plays a significant role in self-driving cars. Tesla, the most popular car manufacturing company is working on self-driving car. It is using unsupervised learning method to train the car models to detect people and objects while driving.

- **Email Spam and Malware Filtering:** Whenever we receive a new email, it is filtered automatically as important, normal, and spam. We always receive an important mail in our inbox with the important symbol and spam emails in our spam box, and the technology behind this is Machine learning. Below are some spam filters used by Gmail:

- Content filter
- Header filter
- General Blacklists filter
- Rules-based filters
- Permission filters

Some machine learning algorithms such as Multi-Layer perceptron**,** Decision tree**,** and Naïve Bayes classifier are used for email spam filtering and malware detection.

- **Virtual Personal Assistant:** We have various virtual personal assistants such as Google assistant, Alexa, Cortana, Siri. As the name suggests, they help us in finding the information using our voice instruction. These assistants can help us in various ways just by our voice instructions such as Play music, call someone, open an email, Scheduling an appointment, etc. These virtual assistants use machine learning algorithms as an important part. These assistants record our voice instructions, send it over the server on a cloud, and decode it using ML algorithms and act accordingly.

- **Online Fraud Detection:** Machine learning is making our online transaction safe and secure by detecting fraud transaction. Whenever we perform some online transaction, there may be various ways that a fraudulent transaction can take place such as fake accounts**,** fake ids, and steal money in the middle of a transaction. So to detect this, Feed Forward Neural network helps us by checking whether it is a genuine transaction or a fraud transaction. For each genuine transaction, the output is converted into some hash values, and these values become the input for the next round. For each genuine transaction, there is a specific pattern which gets change for the fraud transaction hence, it detects it and makes our online transactions more secure.

- **Stock Market Trading**: Machine learning is widely used in stock market trading. In the stock market, there is always a risk of up and downs in shares, so for this machine learning's long short-term memory neural network is used for the prediction of stock market trends.

- **Medical Diagnosis:** In medical science, machine learning is used for diseases diagnoses. With this, medical technology is growing very fast and able to build 3D models that can predict the exact position of lesions in the brain. It helps in finding brain tumours and other brain-related issues.

4. ALGORITHM DESCRIPTION

4.1 Dataset

Data set is garnered 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 patients' clinical data contained likely 40 measurements of vital sign, laboratory, and demographics data. Each file has data separated with pipes in which each row represents a 1 hours' worth of data.

Extremely Imbalance data: The records are extremely imbalanced (More than 97.8% are having 0 sepsis label and 2.2% having sepsis) with the minority class being Sepsis (Shown in Fig 2).

Missing Data: In the data set the percentage of data which is missing high (Shown in Fig 3).This is handled by ignoring the features with more than 80% of missing data.

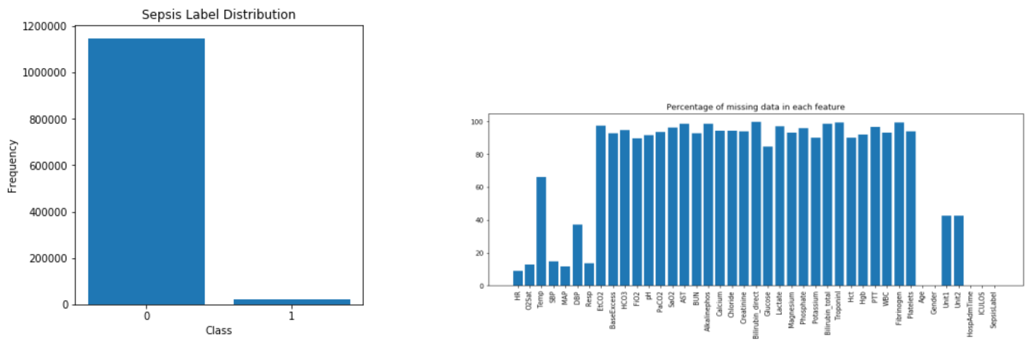


Figure 4.1 Sepsis dataset before re-sampling

4.1.1 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) are the **Labels** for identification.

These are the columns that are in the data set (Shown in fig 3, 4, 5, 6)

Vital signs (columns 1-8)
HR Heart rate (beats per minute)
O2Sat Pulse oximetry (%)
Temp Temperature (Deg C)
SBP Systolic BP (mm Hg)
MAP Mean arterial pressure (mm Hg)
DBP Diastolic BP (mm Hg)
Resp Respiration rate (breaths per minute)
EtCO2 End tidal carbon dioxide (mm Hg)

Figure 4.1.1(a) Vital signs in Dataset

Laboratory values (columns 9-34)
BaseExcess Measure of excess bicarbonate (mmol/L)
HCO3 Bicarbonate (mmol/L)
FiO2 Fraction of inspired oxygen (%)
pH N/A
PaCO2 Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2 Oxygen saturation from arterial blood (%)
AST Aspartate transaminase (IU/L)
BUN Blood urea nitrogen (mg/dL)
Alkalinephos Alkaline phosphatase (IU/L)
Calcium (mg/dL)
Chloride (mmol/L)
Creatinine (mg/dL)
Bilirubin_direct Bilirubin direct (mg/dL)
Glucose Serum glucose (mg/dL)
Lactate Lactic acid (mg/dL)
Magnesium (mmol/dL)
Phosphate (mg/dL)
Potassium (mmol/L)
Bilirubin_total Total bilirubin (mg/dL)
TroponinI Troponin I (ng/mL)
Hct Hematocrit (%)
Hgb Hemoglobin (g/dL)
PTT partial thromboplastin time (seconds)
WBC Leukocyte count (count*10^3/µL)
Fibrinogen (mg/dL)
Platelets (count*10^3/µL)

Figure 4.1.1(b) Laboratory Values in Dataset

Demographics (columns 35-40)
Age Years (100 for patients 90 or above)
Gender Female (0) or Male (1)
Unit1 Administrative identifier for ICU unit (MICU)
Unit2 Administrative identifier for ICU unit (SICU)
HospAdmTime Hours between hospital admit and ICU admit
ICULOS ICU length-of-stay (hours since ICU admit)

Figure 4.1.1(c) Demographics in Dataset

Outcome (column 41)
SepsisLabel: For sepsis patients, SepsisLabel is 1 if $t \geq t_{sepsis}-6$ and 0 if $t < t_{sepsis}-6$. For non-sepsis patients, SepsisLabel is 0.

Figure 4.1.1(d) Outcome in Dataset

4.2 Pre-processing

Dataset has been converted from pipe separated file to comma separated file and with the help of resampling the data have been balanced.

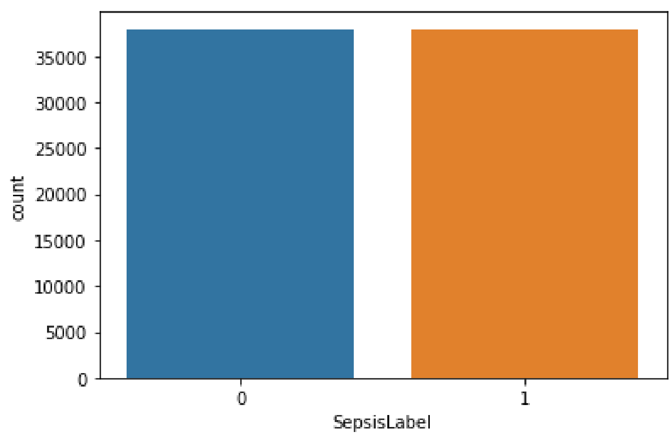


Figure 4.2 Sepsis database after re-sampling

4.3 Feature importance

The features which are important for early prediction of sepsis are selected with help of Xgboost Algorithm (shown in Figure 5), benefit of using this is that after the boosted trees are constructed, it is straightforward to get the importance scores for each of the attributes in the dataset .Generally, it provides a score that indicates how useful each feature in the model. The more an attribute is important it will have a higher score of importance. These are ranked based on the comparison of other attributes in the dataset.

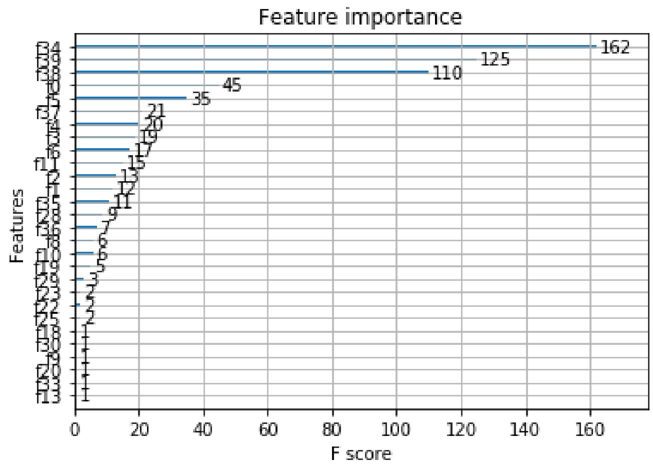


Figure 4.3 Feature Importance Score using XGBoost

4.4 Model Selection

The selection of the model is a crucial factor, as we use machine learning algorithms to forecast the best outcomes.

The supervised approach is used as a preparation for a collection of input and output couples and to test the input and output association pattern. Supervised learning problems are grouped in two issues:

1. Regression analysis: Actual and Constant values in the target or output variable.

2. Classification: Problems not needed for filtering results.

We have a dataset of 40 Dependent variables (parameters or features) and an indie variable is the aim or performance variable, which determines whether or not patients become sepsis. Various classification algorithms are used in this project which are described as follows. This dataset is validated using data mining methods such as the tool for classifying the target groups mentioned above.

Classifiers: A guided learning technique that helps computers to learn from the knowledge. This knowledge was given and then used to identify new findings. The dataset has been analysed using the following classifiers:

1. MLP
2. AdaBoostClassifier
3. Gradient Boosting Classifier
4. GaussianNB
5. LDA
6. QDA

- **MLP Classifier:** MLPClassifier also known as Multi-layer Perceptron classifier which itself suggests a Neural Network. MLPClassifier 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.

ALGORITHM:

1. Take the sepsis dataset and perform re-sampling of the information by re-sampling method so as to balance the dataset.
2. Utilizing Xgboost Algorithm, get the component significance and take out the less significant highlights.
    - 2.1 Use XGBClassifier() and store it as a variable
    - 2.2 fit(X,y) where X and Y are input and output labels respectively
3. Take this Preprocessed dataset isolate it into Training and Testing dataset as X_train,Y_train and X_test,Y_test respectively,
    - 3.1 Train the MLPClassifier model with the Training dataset(X)
        - 3.1.1 MLPClassifier Function i.e., MLPClassifier() with following
            - 3.1.1.1 In this function fixing the following values for the parameters
              hidden_layer_sizes =(,4),
              activation='tanh',
              solver = 'lbfgs',
              max_iter = 5000
            - 3.1.1.2 Putting away the arrival esteem into a variable
        - 3.1.2 Fit the train information into MLPClassifier using fit(X,Y) where X and Y are input and output labels respectively
    - 3.2 Validate the prepared model with Testing dataset
    - 3.3 Print the accuracy

Figure 4.4(a) Algorithm

```
Confusion Matrix :
[[6886  703]
 [ 173 7416]]
Accuracy Score : 0.9422848860192383
Report :
              precision    recall  f1-score   support

           0       0.98      0.91      0.94      7589
           1       0.91      0.98      0.94      7589

    accuracy                           0.94     15178
   macro avg       0.94      0.94      0.94     15178
weighted avg       0.94      0.94      0.94     15178
```

Figure 4.4(b) Confusion Matrix of MLP Classifier

```
===============================
MLPClassifier
****Results****
Accuracy: 94.0506%
Log Loss: 0.17823628759351812
===============================
```

Figure 4.4(c) Results obtained using Multi-Layer Perceptron

The model that is being built using MLP Classifier, 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. 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.

- **Ada-boost or Adaptive Boosting:** It associates multiple classifiers in order to increase the veracity of classifiers and is an iterative ensemble method. This classifier frames a robust classifier by associating multiple below par performing classifiers so that we get a high veracity robust classifier.

```
===============================
AdaBoostClassifier
****Results****
Accuracy: 79.8261%
Log Loss: 0.6731792683873681
===============================
```

Figure 4.4(d) Results obtained using Adaptive Boosting

- **Gradient boosting:** It constructs a prediction model in the form of a collection of weak prediction models, more often than not, decision trees. The model is built in a stage-wise fashion like the other boosting methods, and it hypothesizes them by granting enhancement of arbitrary differentiable loss function.

```
==============================
GradientBoostingClassifier
****Results****
Accuracy: 91.3954%
Log Loss: 0.3111178748731421
==============================
```

Figure 4.4(e) Results obtained using Gradient Boosting Classifier

- **Guassian Naive Bayes:** It is an uncomplicated procedure for building classifiers: models that designate class labels to problem instances, expressed as vectors of factor values, where these class labels are taken from some finite set. There isn't a sole algorithm for instructing such classifiers, but a tribe of algorithms based on a typical principle: all naive Bayes classifiers infer that a particular feature's value is sovereign of any other feature's value, given the class variable.

```
==============================
GaussianNB
****Results****
Accuracy: 57.7678%
Log Loss: 2.120984212868764
```

Figure 4.4(f) Results obtained using Gaussian Naive Bayes

- **Linear Discriminant Analysis (LDA):** It is a technique of dimensionality reduction. As the name entails this technique reduces the total number of dimensions (i.e. variables) in a dataset while confining as much knowledge as possible.

```
==============================
LinearDiscriminantAnalysis
****Results****
Accuracy: 72.4074%
Log Loss: 0.5490759157904654
```

Figure 4.4(g) Results obtained using Linear Discriminant Analysis

- **Quadratic Discriminant Analysis (QDA):** It is a variation of LDA in which a singular covariance matrix is predicted for each and every class of observations. QDA is notably useful if a preceding knowledge exists that individual classes display noticeable covariance. It cannot be used as a technique of dimensionality reduction which is a downside of QDA.

```
==============================
QuadraticDiscriminantAnalysis
****Results****
Accuracy: 50.9224%
Log Loss: 14.830397068643624
==============================
```

Figure 4.4(h) Results obtained using Quadratic Discriminant Analysis

5. SYSTEM ANALYSIS

5.1 System Requirements:

5.1.1 Jupyter Notebook

The IPython Notebook is now known as the Jupyter Notebook. It is an interactive computational environment, in which you can combine code execution, rich text, mathematics, plots and rich media.

The notebook extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results. The IPython notebook combines two components:

A web application: a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output.

Notebook documents: a representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects.

Figure 5.1.1 Jupyter Notebook Logo

5.1.2 Plotting

One major feature of the notebook is the ability to display plots that are the output of running code cells. IPython is designed to work seamlessly with the matplotlib plotting library to provide this functionality.

With matplotlib; it does *not*, however, actually execute any Python import commands, that is, no names are added to the namespace.

If the %matplotlib magic is called without an argument, the output of a plotting command is displayed using the default matplotlibbackend in a separate window. Alternatively, the backend can be explicitly requested using, for example:

```
%matplotlib gtk
```

A particularly interesting backend, provided by IPython, is the inlinebackend. This is available only for the IPython Notebook and the IPython QtConsole. It can be invoked as follows:

```
%matplotlib inline
```

With this backend, the output of plotting commands is displayed *inline* within the notebook, directly below the code cell that produced it. The resulting plots will then also be stored in the notebook document.

Importing .py files

.py files will be imported as a notebook with the same base name, but an .ipynb extension, located in the notebook directory. The notebook created will have just one cell, which will contain all the code in the .py file. You can later manually partition this into individual cells using the Edit | Split Cell menu option, or the Ctrl-m - keyboard shortcut.

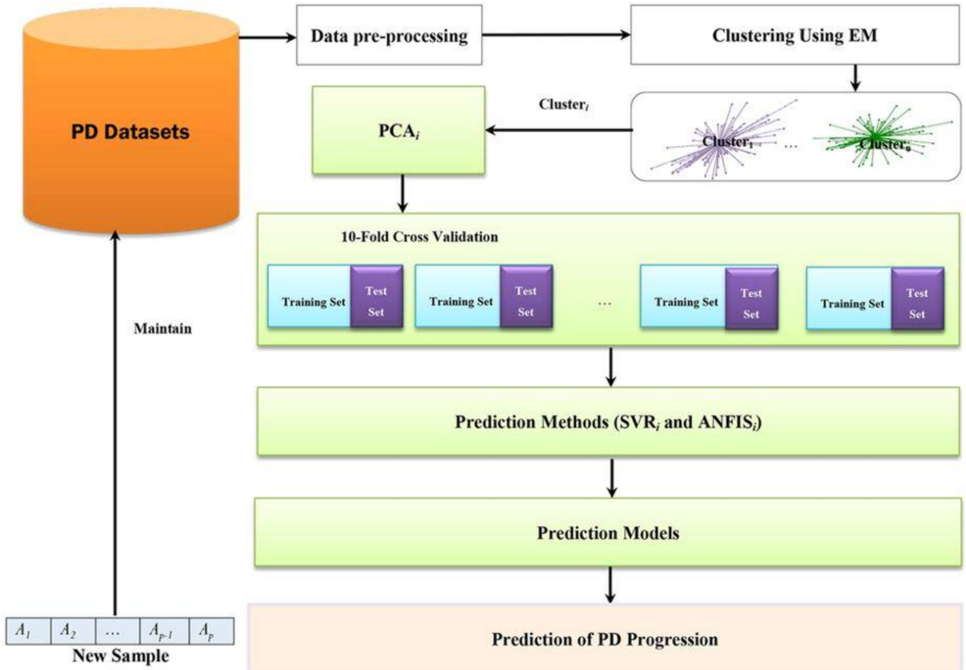


Figure 5.1.2 Training and Testing data set

6. SYSTEM DESIGN

6.1 UML Diagrams Introduction:

UML is a standard language for specifying, visualizing, constructing, and documenting the artefacts of software systems. 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 as well 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.

6.2 Activity Diagram:

6.2.1 Definition:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modelling Language, activity diagrams are intended to model both computational and organizational processes. Activity diagrams show the overall flow of control.

6.2.2 Activity Diagram for Sepsis detection

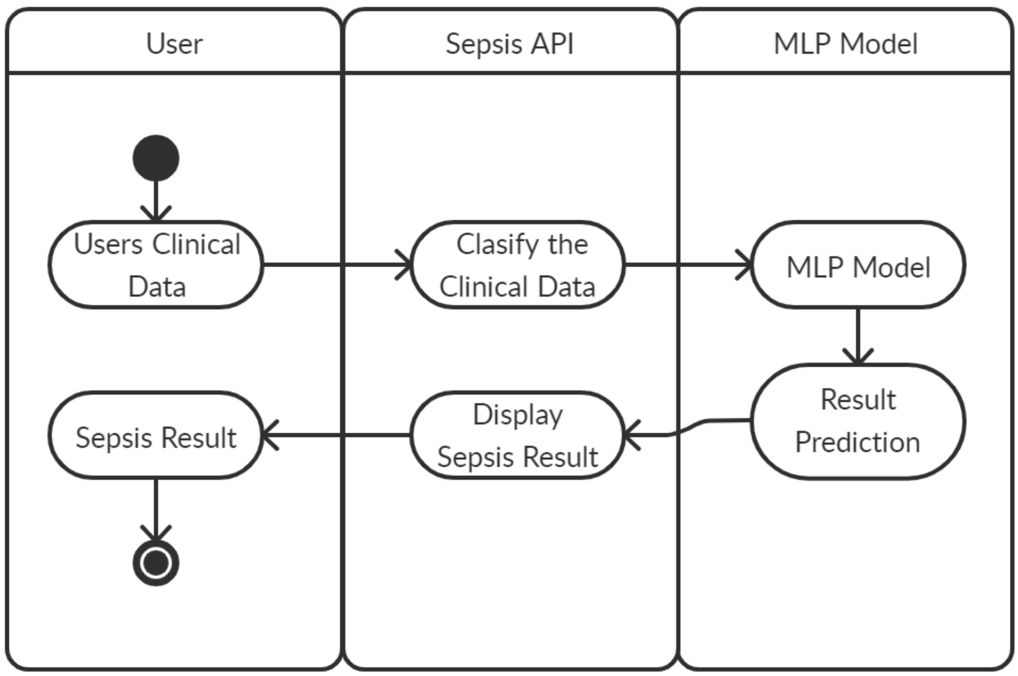


Figure 6.2.2 Activity diagram for Sepsis detection using machine learning algorithms

6.3 Class Diagram:

6.3.1 Definition:

A **class diagram** in the Unified Modelling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects.

6.3.2 Class Diagram for Sepsis detection

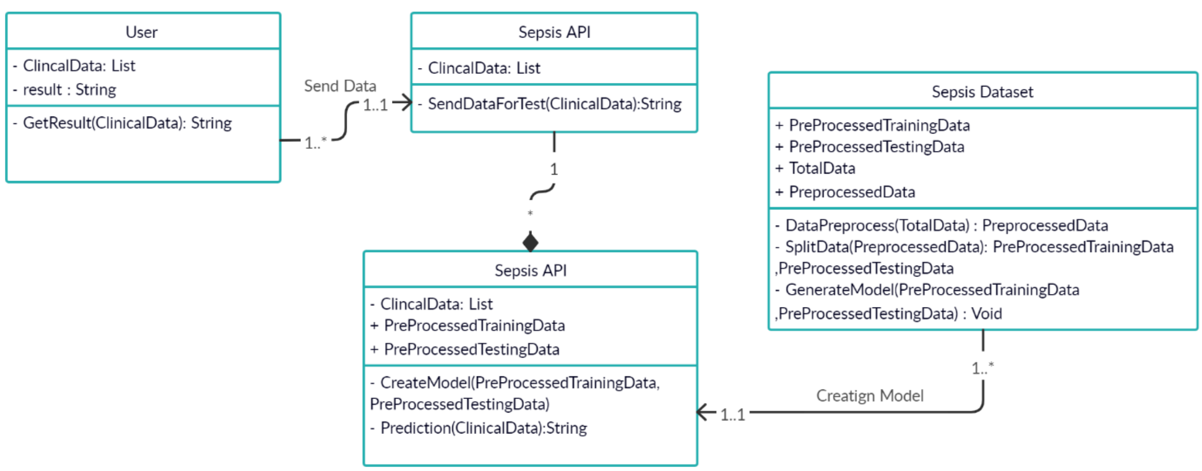


Figure 6.3.2 Class diagram for Sepsis detection using machine learning algorithms

6.4 Use case Diagram:

6.4.1 Definition

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. A use case represents a distinct functionality of a system, a component, a package, or a class.

6.4.2 Use case Diagram for Sepsis detection

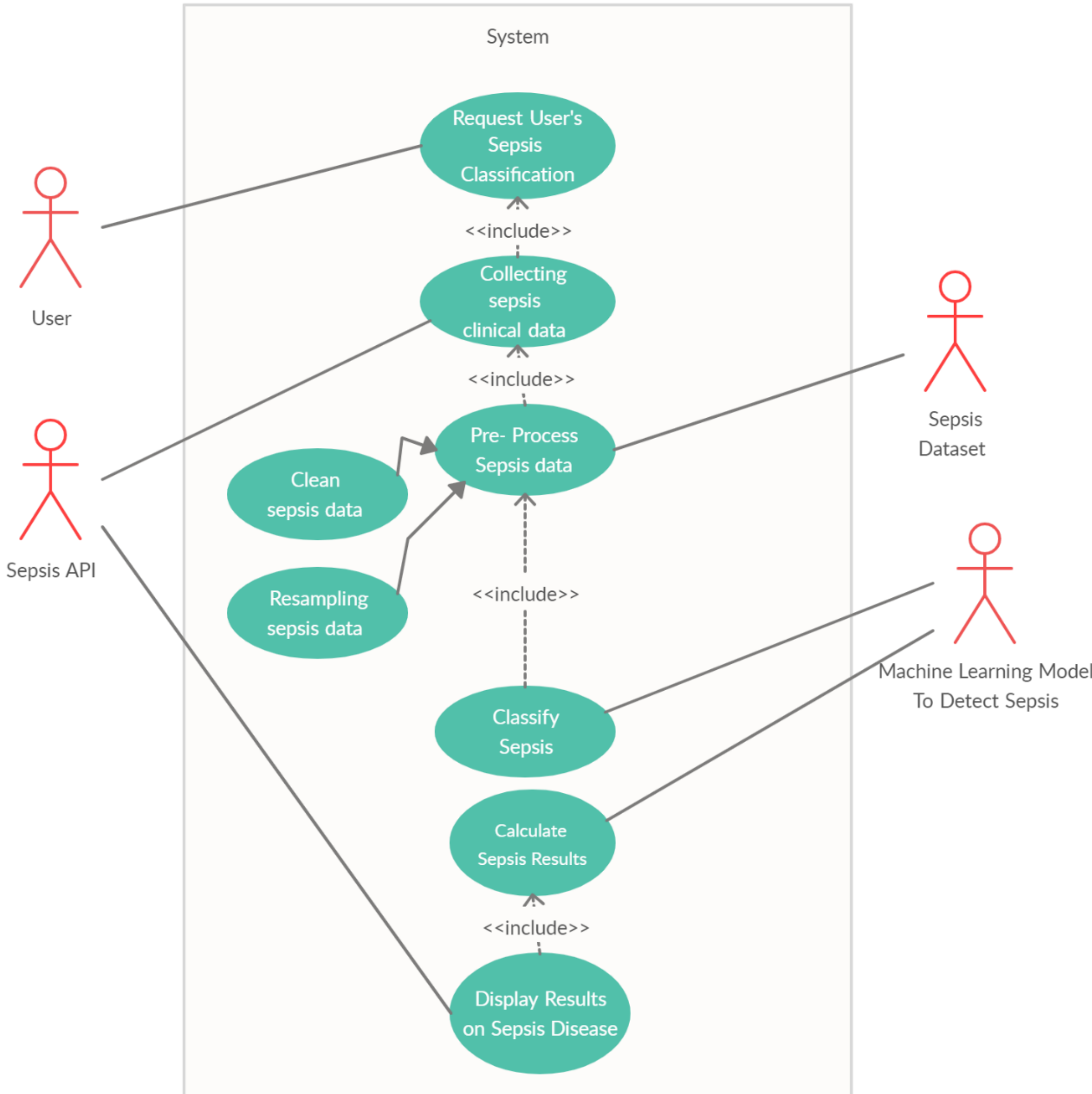


Figure 6.4.2 Use case diagram for Sepsis detection using machine learning algorithms

6.5 Sequence Diagram:

6.5.1 Definition:

A **sequence diagram** simply depicts interaction between objects in a sequential order i.e. the order in which these interactions take place. We can also use the terms event diagrams or event scenarios to refer to a sequence diagram. Sequence diagrams describe how and in what order the objects in a system function.

6.5.2 Sequence Diagram for Sepsis detection

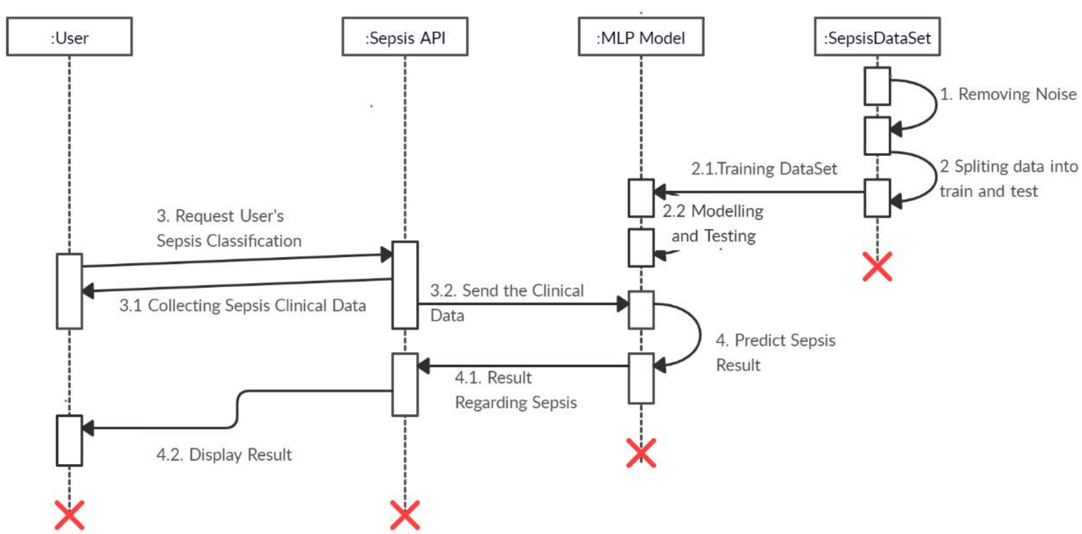


Figure 6.5.2 Sequence diagram for Sepsis detection using machine learning algorithms

7. IMPLEMENTATION

7.1 Coding:

Flow Chart of Implementation

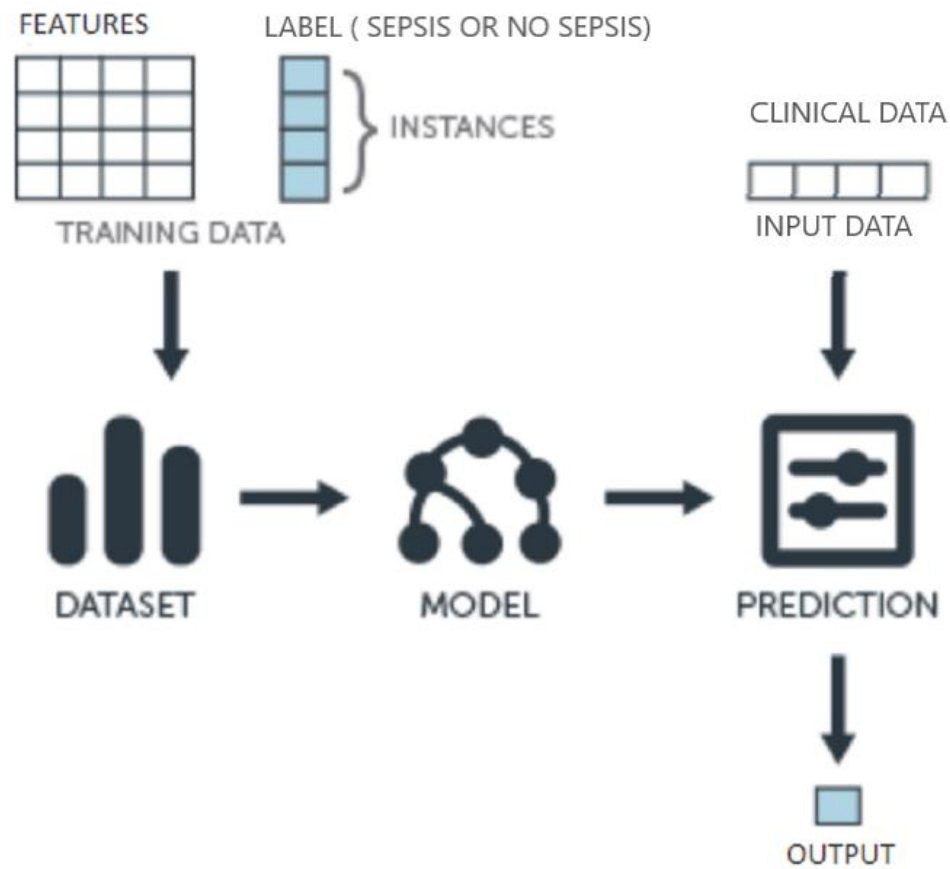


Figure 7.1 Flow Chart of Implementation

We can define the machine learning work flow in 5 stages.

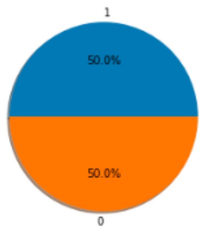
- Gathering data
- Data pre-processing
- Feature Importance
- Researching the model that will be best for the type of data
- Training and testing the model

- Evaluation

7.1.1 Code for MLP Classifier Algorithm

```
In [2]: import matplotlib.pyplot as plt
        import seaborn as sns
        import pandas as pd
        from sklearn.preprocessing import LabelEncoder
        from sklearn.model_selection import train_test_split
        from sklearn.neural_network import MLPClassifier

In [3]: dataset = pd.read_csv("C:/Users/hcyen/SCA/sepsis.csv")

In [4]: from sklearn.utils import resample
        df_majority = dataset[dataset.SepsisLabel==0]
        df_minority = dataset[dataset.SepsisLabel==1]

In [5]: df_minority_upsampled = resample(df_minority,
                                         replace=True,     # sample with replacement
                                         n_samples=37945,    # to match majority class
                                         random_state=123) # reproducible results

In [6]: df_upsampled = pd.concat([df_majority, df_minority_upsampled])
        df_upsampled.SepsisLabel.value_counts()

Out[6]: 1    37945
        0    37945
        Name: SepsisLabel, dtype: int64
```

```
In [7]: plt.pie(df_upsampled['SepsisLabel'].value_counts(), labels=['1','0'], autopct='%1.1f%%', shadow=True)
        plt.show()
```




```
In [8]: X =df_upsampled[df_upsampled.columns[0:40]].values

In [9]: Y = df_upsampled[df_upsampled.columns[40:]].values

In [10]: print("sca dimensions : {}".format(df_upsampled.shape))

         sca dimensions : (75890, 41)

In [11]: print("sca dimensions : {}".format(X.shape))

         sca dimensions : (75890, 40)

In [12]: print("sca dimensions : {}".format(Y.shape))

         sca dimensions : (75890, 1)
```

```
In [15]: labelencoder_Y = LabelEncoder()
         Y = labelencoder_Y.fit_transform(Y)

C:\Users\hcyen\Anaconda3\lib\site-packages\sklearn\preprocessing\label.py:235: DataConversionWarning: A column-vector y was pas
sed when a 1d array was expected. Please change the shape of y to (n_samples, ), for example using ravel().
  y = column_or_1d(y, warn=True)
```

```
In [16]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.20, random_state=0)
         print("Training data dimensions :{}".format(X_train.shape))
         print("Testing data dimensions :{}".format(X_test.shape))

Training data dimensions :(60712, 40)
Testing data dimensions :(15178, 40)
```

```
In [17]: clf = 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=5000,
             learning_rate_init=1e-5,
             tol=1e-4,
         )
```

```
In [18]: clf.fit(X_train, Y_train)

Out[18]: MLPClassifier(activation='tanh', alpha=0.0001, batch_size='auto', beta_1=0.9,
                beta_2=0.999, early_stopping=False, epsilon=1e-08,
                hidden_layer_sizes=(40, 10, 10, 10, 10, 2),
                learning_rate='constant', learning_rate_init=1e-05, max_iter=5000,
                momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True,
                power_t=0.5, random_state=1, shuffle=True, solver='lbfgs',
                tol=0.0001, validation_fraction=0.1, verbose=False,
                warm_start=False)
```

```
In [19]: predicted = clf.predict(X_test)
         idx = 0
         true = 0
         false = 0
         for i in X_test:
             # Printing the predicted values for each test case
             #print ("%d - %d" %(predicted[idx], Y_test[idx]))
             #print(predicted[idx])
             # Calculating the number of correct predictions and wrong predictions
             if predicted[idx] == Y_test[idx]:
                 true += 1
             else:
                 false += 1
             idx += 1
         print("Number of Correct Predictions:")
         # Printing the number of correct predictions
         print(true)
         print("Number of Wrong Predictions:")
         # Printing the number of wrong predictions
         print(false)
         # Accuracy is calculated
         print("Accuracy of the MLPClassifier:")
         accuracy = (true / (true + false)) * 100
         print(accuracy)

Number of Correct Predictions:
14285
Number of Wrong Predictions:
893
Accuracy of the MLPClassifier:
94.11648438529451
```

```
Confusion Matrix :
[[6886  703]
 [ 173 7416]]
Accuracy Score : 0.9422848860192383
Report :
              precision    recall  f1-score   support

           0       0.98      0.91      0.94      7589
           1       0.91      0.98      0.94      7589

    accuracy                           0.94     15178
   macro avg       0.94      0.94      0.94     15178
weighted avg       0.94      0.94      0.94     15178
```

Figure 7.1.1 MLP Classifier Algorithm Code

TESTING OF IMPLEMENTATION OF MLP ALGORITHM

```
predicted = clf.predict(X_test)
idx = 0
true = 0
false = 0
print("\npredicted result - original result")
for i in X_test:
    # Printing the predicted values for each test case
    if(predicted[idx]==0):
        a="NoSepsis"
    else:
        a="Sepsis  "
    if(Y_test[idx]==0):
        b="NoSepsis"
    else:
        b="Sepsis"
    print(a ,"         ", b)
```

```
predicted result - original result
NoSepsis            NoSepsis
NoSepsis            NoSepsis
Sepsis              Sepsis
Sepsis              NoSepsis
Sepsis              Sepsis
NoSepsis            NoSepsis
Sepsis              Sepsis
NoSepsis            NoSepsis
Sepsis              Sepsis
NoSepsis            NoSepsis
```

Figure 7.1.2 Screenshot of Testing of Implementation of MLP Classifier Algorithm

8. COMPARISION OF DIFFERENT MODELS

8.1 MLP Classifier: MLPClassifier also known as Multi-layer Perceptron classifier which itself suggests a Neural Network. MLPClassifier 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

A) ***Real-world application***

The neural network literature is full of pattern recognition applications. Typically one takes pixelated image values as the network input and that maps via layers of hidden units to a set of outputs corresponding to possible classifications of the image.

An early, but typical, example by Le Cun et al. (1989) was designed to recognise handwritten ZIP codes (i.e. numerical postal codes). The inputs consisted of a 16 × 16 array representing pixelated images of hand-written digits scaled to a standard size, and these fed through three layers of hidden units to ten output units which each corresponded to one of the digits 0–9. The first hidden layer contained 12 feature detectors (8 × 8), and the second contained 12 feature detectors (4 × 4). Each unit in each detector had a 5 × 5 receptive field in the earlier layer, and hard weight sharing was used to ensure that they all detected the same feature in different parts of the retina. The third hidden layer had 30 units fully connected to the second hidden layer and the outputs.

The network was trained on 7300 digits with ~1% errors and tested on 2000 digits with ~5% errors. Pruning by Optimal Brain Damage improved the performance further.

B) Strengths of the model

Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.

They yield the required decision function directly via training.

MLP/Neural networks do not make any assumption regarding the underlying probability density functions or other probabilistic information about the pattern classes under consideration in comparison to other probability based models.

C) Weakness of the model:

Hardware dependence: Artificial neural networks require processors with parallel processing power, in accordance with their structure. For this reason, the realization of the equipment is dependent.

Determination of proper network structure: There is no specific rule for determining the structure of artificial neural networks. Appropriate network structure is achieved through experience and trial and error.

The duration of the network is unknown: The network is reduced to a certain value of the error on the sample means that the training has been completed. This value does not give us optimum results.

8.2 Ada-boost or Adaptive Boosting:

It associates multiple classifiers in order to increase the veracity of classifiers and is an iterative ensemble method. This classifier frames a robust classifier by associating multiple below par performing classifiers so that we get a high veracity robust classifier.

A) Real-world application

AdaBoost can be used to solve a variety of real-world problems, such as predicting customer churn and classifying the types of topics customers are talking/calling about. The algorithm is heavily utilised for solving classification problems, given its relative ease of implementation in languages such as R and Python

B) Strengths of the model

Very good use of weak classifiers for cascading;

Different classification algorithms can be used as weak classifiers;

AdaBoost has a high degree of precision;

Relative to the bagging algorithm and Random Forest Algorithm, AdaBoost fully considers the weight of each classifier;

C) Weakness of the model

The number of AdaBoost iterations is also a poorly set number of weak classifiers, which can be determined using cross-validation;

Data imbalance leads to a decrease in classification accuracy;

Training is time consuming, and it is best to cut the point at each reselection of the current classifier;

8.3 Gradient Boosting: GBT build trees one at a time, where each new tree helps to correct errors made by previously trained tree.

A) Real-world application

A great application of GBM is anomaly detection in supervised learning settings where data is often highly unbalanced such as DNA sequences, credit card transactions or cyber security.

Reference presents a more specific application in this context, supervised anomaly detection task with a learning to rank approach. Learning to rank means the application of machine learning in the construction of ranking models for information retrieval systems. This results in finding the anomalies with the highest precision without giving too many genuine examples to the experts. According to this manuscript, gradient boosting has shown to be a powerful method on real-life datasets to address learning to rank problems due to its two main features:

It performs the optimization in function space (rather than in parameter space) which makes the use of custom loss functions much easier.

Boosting focuses step by step on difficult examples that give a nice strategy to deal with unbalanced datasets by strengthening the impact of the positive class.

B) Strengths of the model

Often provides predictive accuracy that cannot be beat.

Lots of flexibility - can optimize on different loss functions and provides several hyper parameter tuning options that make the function fit very flexible.

No data pre-processing required - often works great with categorical and numerical values as is.

Handles missing data - imputation not required.

C) Weaknesses of the model

GBMs will continue improving to minimize all errors. This can overemphasize outliers and cause over fitting. Must use cross-validation to neutralize.

Computationally expensive - GBMs often require many trees (>1000) which can be time and memory exhaustive.

The high flexibility results in many parameters that interact and influence heavily the behaviour of the approach (number of iterations, tree depth, regularization parameters, etc.). This requires a large grid search during tuning.

Less interpretable although this is easily addressed with various tools (variable importance, partial dependence plots, LIME, etc.).

8.4 Gaussian Naive Bayes: It is an uncomplicated procedure for building classifier models that designate class labels to problem instances, expressed as vectors of factor values, where these class labels are taken from some finite set. There isn't a sole algorithm for instructing such classifiers, but a tribe of algorithms based on a typical principle: all naive Bayes classifiers infer that a particular feature's value is sovereign of any other feature's value, given the class variable.

A) Real-world application

Text classification/ Spam Filtering/ Sentiment Analysis: Naive Bayes classifiers are mostly used in text classification (due to their better results in multi-class problems and independence rule) have a higher success rate as compared to other algorithms. As a result, it is widely used in Spam filtering (identify spam e-mail) and Sentiment Analysis (in social media analysis, to identify positive and negative customer sentiments)

B) Strengths of the model

The naive Bayesian model originated from classical mathematical theory and has a solid mathematical foundation and stable classification efficiency.

It has a higher speed for large numbers of training and queries. Even with very large training sets, there is usually only a relatively small number of features for each project, and the training and classification of the project is only a mathematical operation of the feature probability.

It works well for small-scale data, can handle multi-category tasks, and is suitable for incremental training (that is, it can train new samples in real time).

Less sensitive to missing data, the algorithm is also relatively simple, often used for text classification.

Naïve Bayes explains the results easily.

C) Weakness of the model

Need to calculate the prior probability.

There is an error rate in the classification decision.

Very sensitive to the form of input data.

The assumption of sample attribute independence is used, so if the sample attributes are related, the effect is not good.

8.5 Linear Discriminant Analysis (LDA): It is a technique of dimensionality reduction. As the name entails this technique reduces the total number of dimensions (i.e. variables) in a dataset while confining as much knowledge as possible.

A) Real-world application

Face Recognition: In the field of Computer Vision, face recognition is a very popular application in which each face is represented by a very large number of pixel values. Linear Discriminant Analysis (LDA) is used here to reduce the number of features to a more manageable number before the process of classification. Each of the new dimensions generated is a linear combination of pixel values, which form a template. The linear combinations obtained using Fisher's linear discriminant are called Fisher faces.

Medical: In this field, Linear Discriminant Analysis (LDA) is used to classify the patient disease state as mild, moderate or severe based upon the patient's various parameters and the medical treatment he is going through. This helps the doctors to intensify or reduce the pace of their treatment.

Customer Identification: Suppose we want to identify the type of customers which are most likely to buy a particular product in a shopping mall. By doing a simple question and answers survey, we can gather all the features of the customers. Here, Linear Discriminant Analysis will help us to identify and select the features which can describe the characteristics of the group of customers that are most likely to buy that particular product in the shopping mall.

B) Strengths of the model

LDA is supervised, which *can* (but doesn't always) improve the predictive performance of the extracted features. Furthermore, LDA offers variations (i.e. quadratic LDA) to tackle specific roadblocks.

C) Weaknesses of the model

The main limitation of the Linear Discriminant Analysis is that the new features are not easily interpretable, and you must still manually set or tune the number of components to keep. LDA also requires labelled data, which makes it more situational.

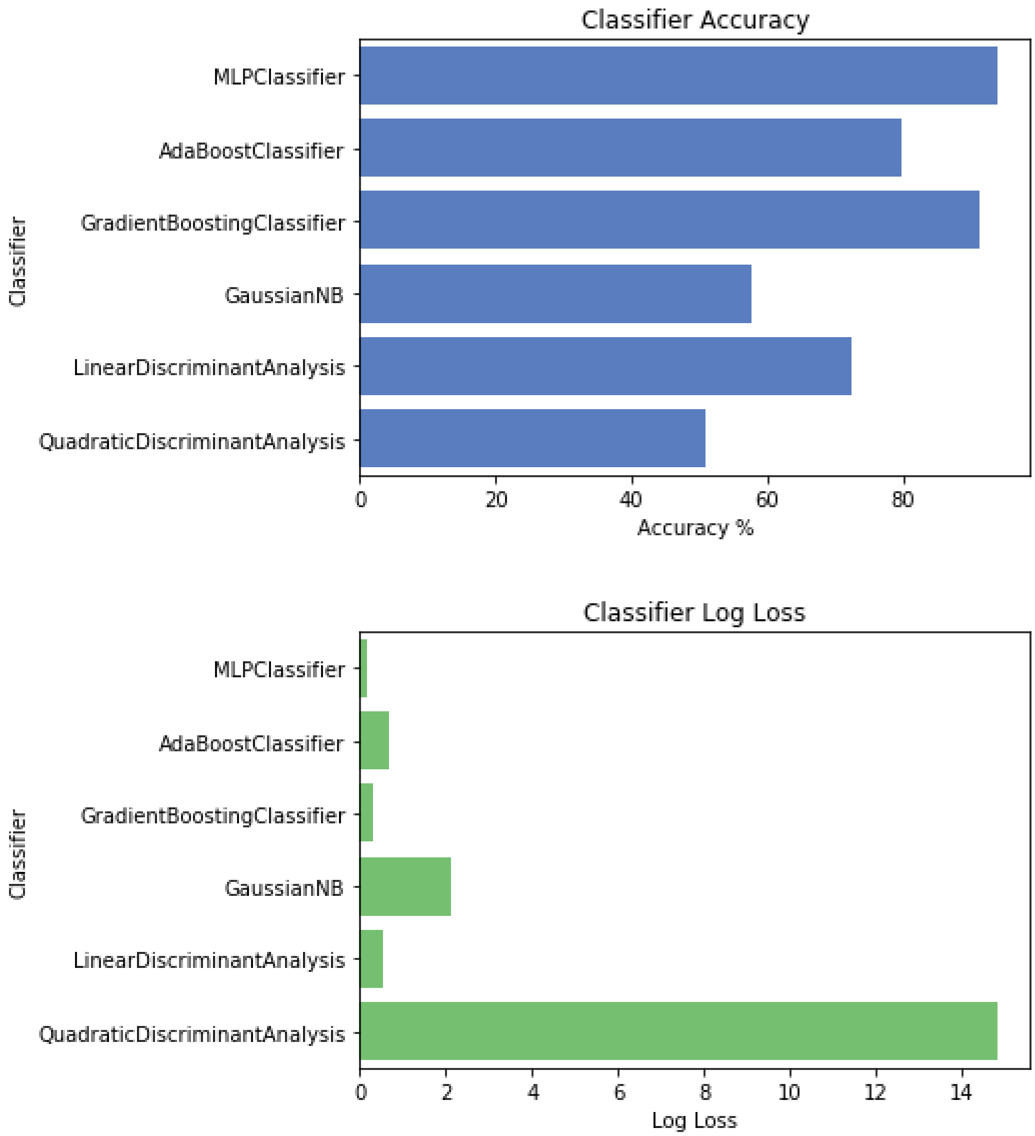


Figure 8.2 Graph of Accuracy and Log Loss Classifier

The results show that the classification algorithms like MLP Classifier, AdaBoostClassifier, GradientBoostingClassifier, GaussianNB, LDA, and QDA gives the accuracy of prediction like 94%, 71.8%, 91.3%, 57.7%, 72.4% and 50.9% respectively. Finally the results showed that the Multi-Layer Perceptron (MLP) gives the best accuracy in identification of sepsis at early stages with help of clinical data available.

9. CONCLUSION AND FUTURE SCOPE

9.1 Conclusion

Sepsis is a hazardous condition brought about 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 paper has presented a description about Sepsis and its history in the international scenarios and in the national scenario (in the context of India). The symptoms of the disease, signs, complications, and treatment for the disease are presented. This paper also presents the detection of this disease at early stages with higher accuracy using various classifiers mainly on using MLP classifiers in order to develop good prediction models which helps in avoiding long time-taking lab results.

9.2 Future Scope

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.

BIBILIOGRAPHY

References

[1] Nachimuthu S.K., Huag P.J. Early Detection of Sepsis in the Emergency Department using Dynamic Bayesian Networks; Proceedings of the 20 12 AMIA Annual Symposium; Chicago, IL, USA. 3–7 November 2012; pp. 653–662. [PMC free article] [PubMed] [Google Scholar]

[2] Sutherland, A., Thomas, M., Brandon, R.A. *et al.* Development and validation of a novel molecular biomarker diagnostic test for the early detection of sepsis. *Crit Care* **15,** R149 (2011). https://doi.org/10.1186/cc10274

[3] Karen K. Giuliano; Physiological Monitoring for Critically Ill Patients: Testing a Predictive Model for the Early Detection of Sepsis. *Am J Crit Care* 1 March 2007; 16 (2): 122–130. doi: https://doi.org/10.4037/ajcc2007.16.2.122

[4] Anderson, Stephen J., Douglas J. Haney, and Cory A. Waters. "Early detection of sepsis." U.S. Patent No. 7,465,555. 16 Dec. 2008.

[5] Fairchild, Karen D. "Predictive monitoring for early detection of sepsis in neonatal ICU patients." *Current opinion in pediatrics* 25.2 (2013): 172-179.

[6] 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," *2019 Computing in Cardiology (CinC)*, Singapore, Singapore, 2019, pp. Page 1-Page 4.

[7] Kam, Hye Jin, and Ha Young Kim. "Learning representations for the early detection of sepsis with deep neural networks." *Computers in biology and medicine* 89 (2017): 248-255.

[8] Taylor, R. Andrew, et al. "Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data–driven, machine learning approach." *Academic emergency medicine* 23.3 (2016): 269-278.

[9] Fagerström, J., Bång, M., Wilhelms, D. *et al.* LiSep LSTM: A Machine Learning Algorithm for Early Detection of Septic Shock. *Sci Rep* **9,** 15132 (2019). https://doi.org/10.1038/s41598-019-51219-4

[10] Shimabukuro DW, Barton CW, Feldman MD, *et al*
Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay: a randomised clinical trial *BMJ Open Respiratory Research* 2017;**4:**e000234. doi: 10.1136/bmjresp-2017-000234

[11] Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L, Shimabukuro D, Chettipally U, Feldman MD, Barton C, Wales DJ, Das R Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach JMIR Med Inform 2016;4(3):e28
DOI: 10.2196/medinform.5909

[12] Tang, Collin HH, Andrey V. Savkin, and Paul M. Middleton. "Non-invasive Sepsis Patient Classification using Least Squares Support Vector Machine." *BIOSIGNALS*. 2009.
[13] Lauritsen, Simon Meyer, et al. "Early detection of sepsis utilizing deep learning on electronic health record event sequences." *Artificial Intelligence in Medicine* (2020): 101820.
[14] Islam, Md Mohaimenul, et al. "Prediction of sepsis patients using machine learning approach: a meta-analysis." *Computer methods and programs in biomedicine* 170 (2019): 1-9.
[15] P. Hsu and C. Holtz, "A Comparison of Machine Learning Tools for Early Prediction of Sepsis from ICU Data," 2019 Computing in Cardiology (CinC), Singapore, Singapore, 2019, pp. Page 1-Page 4.
[16] https://www.healthline.com/health/sepsis
[17] https://physionet.org/content/challenge-2019/1.0.0/