

Using Machine Learning Algorithms to predict Sepsis and its stages in ICU patients



By

Nimrah Ghias

Fall-2019-MSBI-00000318181

Batch

2019-2021

Supervised by:

Dr. Mehak Rafiq

Research Centre for Modelling and Simulation (RCMS)

National University of Science & Technology (NUST)

Islamabad, Pakistan.

Submission Date

January, 2022

Certificate of Originality

I hereby declare that the results presented in this research work titled as “Using machine learning algorithms to predict sepsis and its stages in ICU patients” are generated by myself. Moreover, none of its contents are plagiarized nor set forth for any kind of evaluation or higher education purposes. I have acknowledged/referenced all the literary content used for support in this research work.

Nimrah Ghias

00000318181

Dedication

I dedicate this research project report to ALLAH (SWT) and my parents, teachers, siblings, friends for their endless support, love, encouragement and invaluable trust in me. All these people are the precious gift of ALLAH (SWT).

Acknowledgment

All gratitude is to the most “Gracious”, the most “Merciful” Allah Almighty, who guided and aided me to bring-forth this report and respect for Holy Prophet Hazrat MUHAMMAD (ﷺ) whose teachings are complete guidance for humanity. I am greatly thankful to Dr. Mehak Rafiq, who guided me towards the success, and it was an amazing experience to work under her supervision. Concisely and precisely, I feel deeply indebted to the company of my beloved Father Ghias Ahmad, Mother Zahida Majeed and siblings Aqsa Sajjad, Faisal Sajjad, Iqra Ghias and Hira Ghias who were encouraging, supportive and heartwarming all the time. Besides my advisor, I would like to pay sincere thanks to my co-supervisor Mr. Shan-ul-Haq, my GEC members Dr. Muhammad Tariq Saeed and Dr. Maria Shabir and my fellows Ateeq ur Rehman, Mohsin Shehzad Khan, Zunera Jamal, Rida Ayub, Huzaifa Arshad, Haseeb Khan, Mehar Masood, Noor Khan, Tayyaba Alvi, Farhan Bashir, Farhana, Mohsin Ahmad, Ameer Ghaznavi for being helping and supportive throughout my project.

List of Figures

Figure 1: Pathophysiology of Sepsis	4
Figure 2: Steps of Methodology	22
Figure 3: Data Description	23
Figure 4: Percentage of Missing Values in Training Set A	35
Figure 5: Percentage of Missing Values in Training Set B	36
Figure 6: Percentage of Missing Values in Combined Dataset A and B	36
Figure 7: Histogram of Training Set A columns	38
Figure 8: Histogram of Training Set B columns	39
Figure 9: Histogram of Combined Training Set A and B	40
Figure 10: Correlation Matrix of Training Set A	40
Figure 11: Correlation Matrix of Training Set B	41
Figure 12: Correlation Matrix of Combine Training Set A and B	42
Figure 13: Gender Analysis of Training Set A	43
Figure 14: Gender Analysis of Training Set B	44
Figure 15: Correlation matrix of Combined Training Set A and B	45
Figure 16: Age Analysis of Training Set A	46
Figure 17: Gender Analysis of Training Set B	46
Figure 18: Gender Analysis of Combined Training Set A and B	47
Figure 19: Number of Septic and Non-Septic patients in Training Set A	48
Figure 20: Number of Septic and Non-Septic patients in Training Set B	49
Figure 21: Number of Septic and Non-Septic Patients in Combined Dataset A and B	49
Figure 22: Confusion Matrix of Xgboost in Training Set A	50
Figure 23: Confusion matrix of LightGBM in Training Set A	51
Figure 24: Confusion matrix of Random Forest in Training Set A	51
Figure 25: Confusion Matrix of Xgboost in Training Set B	52
Figure 26: Confusion Matrix of LightGBM in Training Set B	52
Figure 27: Confusion Matrix of Random Forest in Training Set B	53
Figure 28: Confusion Matrix of Xgboost in Combined Training Set A and B	53
Figure 29: Confusion Matrix of LightGBM in Combined Training Set A and B	54
Figure 30: Confusion Matrix of Random Forest in Combined Training Set A and B	54
Figure 31: Classification Report of Xgboost in Training Set A	55
Figure 32: Classification Report of LightGBM in Training Set A	55
Figure 33: Classification Report of RandomForest in Training Set A	56
Figure 34: Classification Report of Xgboost in Training Set B	56
Figure 35: Classification Report of LightGBM in Training Set B	57
Figure 36: Classification Report of RandomForest in Training Set B	57
Figure 37: Classification Report of Xgboost in Combined Training Set A and B	58
Figure 38: Classification Report of LightGBM in Combined Training Set A and B	58
Figure 39: Classification Report of RandomForest in Combined Training Set A and B	59
Figure 40: ROC curve of Training Set A	60
Figure 41: ROC curve of Training Set B	61
Figure 42: ROC curve of Combined Training Set A and B	62

Contents

Certificate of Originality.....	i
Dedication.....	ii
Acknowledgment.....	iii
List of Figures.....	iv
List of Abbreviations.....	viii
Abstract.....	ix
Chapter 1.....	1
Introduction:.....	1
1.1 Sepsis.....	1
1.2 Pathophysiology of sepsis:.....	2
1.3 Multiple Organ Dysfunction Syndrome:.....	4
1.4 Assessment of Sepsis.....	5
1.4.1 SIRS:.....	5
1.4.2 qSOFA:.....	6
1.4.3 NEWS.....	6
1.5 Machine Learning:.....	7
Chapter 2.....	9
Literature Review:.....	9
Chapter 3.....	22
Methodology:.....	22
3.1 Data Collection:.....	22
3.2 Tools Used:.....	23
3.3 Data Preprocessing:.....	24
3.3.1 Case Deletion:.....	24
3.3.2 Mean Imputation:.....	25
3.3.3 Median Imputation:.....	25
3.3.4 Pre and Next Imputation:.....	26
3.3.5 Mode Imputation:.....	26
3.3.6 0 imputation:.....	26

3.3.7 Missforest Imputation:	26
3.4 Feature Selection:.....	27
3.5 Correlation Analysis:	27
3.6 Statistical Analysis of Data:.....	28
3.6.1 Gender Analysis:.....	28
3.6.2 Sepsis Label 0 and 1:	29
3.6.3 Age Analysis:.....	29
3.7 Train/ Test Split:	29
3.8 Smote Analysis:	30
3.9 Machine learning Algorithms:	30
3.10 Cross Validation:	33
3.10.1 Hold Out Method:	33
3.10.2 K fold cross validation:	34
3.10.3 Leave One Out Cross Validation:	34
Chapter 4.....	35
Results:.....	35
4.1 Percentage of Missing Values in Training Sets:	35
4.2 Histogram of Imputed Values in every column of Training Sets	37
4.3 Correlation and Statistical Analysis:.....	40
4.3.1 Correlation Matrix:	40
Training Set A:.....	40
Training Set B:.....	41
Training Set AB:	42
4.3.2 Gender Analysis:.....	42
4.3.3 Age Analysis:.....	45
4.3.4 Septic and Non Septic Patients:	47
4.4 Confusion Matrix after Smote Analysis:	50
4.5 Classification Report:.....	55
4.6 Roc Curve:	60
Training Set A:.....	60
Chapter 5.....	63
Discussion:	63
Conclusion:.....	65

References: 66

List of Abbreviations

RIG-I	Retinoic acid inducible gene I
PRR	Pattern recognition receptors
NOD	Nucleotide-oligomerization domain
ATP	Adenine Tri-Phosphate
DAMPs	Danger associated molecular patterns
NET	Neutrophil extracellular traps
TLR	Toll Like Receptor
PMNs	Polymorphonuclear
IL-1	Interleukin-1
TNFα	Tumor necrosis factor alpha
MODS	Multiple organ dysfunction syndrome
PaO₂	Partial pressure of arterial oxygen
AUC	Area Under Curve

Abstract

Sepsis is blood poisoning disease that occurs when body shows dysregulated host response to an infection and cause organ failure or tissue damage which may increase the mortality rate in ICU patients. As it becomes major health problem, the hospital cost for treatment of sepsis is increasing every year. Different methods have been developed to monitor sepsis electronically, but it is necessary to predict sepsis as soon as possible before clinical reports or traditional methods, because delayed in treatment can increase the risk of mortality with every single hour. For the early detection of sepsis, specifically in ICU patients, different machine learning models i.e., Linear learner, Multilayer perceptron neural networks, Random Forest, Lightgbm and Xgboost has trained on the data set proposed by Physio Net/ Computing in Cardiology Challenge in 2019. This study shows that Machine learning algorithms can accurately predict sepsis at the admission time of patient in ICU by using six vital signs extracted from patient records over the age of 18 years. After comparative analysis of machine learning models, Xgboost, Randomforest and Lightgbm model achieved a highest accuracy of under the range of 0.89-0.96, precision of 0.90-0.96, and recall 0.78-0.96 under the precision-recall curve on the publicly available data. Early prediction of sepsis can help clinicians to implement supportive treatments and reduce the mortality rate as well as healthcare expenses.

Chapter 1

Introduction:

1.1 Sepsis

There was a lot of confusion to define systematic response syndrome to infection before 1992. But the consensus meeting has confirmed the definition, as sepsis is systematic inflammatory response syndrome after the confirmation of bacterial infection. It is also considered as life threatening disease because it causes organ dysfunction that occurs when body shows extreme response to an infection. Many studies have validated that other two types of sepsis i.e. (severe sepsis and septic shock) are the biomarkers to increase the mortality rate. In 2001, the other consensus meeting proposed that sepsis should be defined on the basis of biomarkers (O'Brien et al., 2007). Severe sepsis is linked with tissue hypoperfusion (oliguria, elevated lactate) and organ dysfunction (coagulopathy). It can be measured by parameters i.e., lactic acid > 2.0mmol/L, SBP < 90mmHg, creatinine 0.5mg/dL, Map < 65mmHg, $100 \times 10^9/L$ etc. while Septic shock is distributive shock that can be defined as sepsis which has cellular, metabolic and circulatory abnormalities that cause higher risk of death than sepsis alone. It includes the patients who fill the criteria of sepsis and needs a vasopressor to balance mean arterial pressure (MAP \geq 65mmHg) and lactate > 2mmol/L. The measurements for the detection of septic shock are

- SBP < 90mmHg,
- MAP < 65mmHg
- lactic acid > 3.9mmol/L etc.

The septic response involves the complicated biological events i.e., anti-inflammatory response, abnormality in blood circulations, cellular reactions etc. It is difficult to diagnose

the sepsis due to its complex events and undefined symptoms. Therefore, the early detection of sepsis is necessary for the specific treatment at suitable time. That is why, biomarkers are very important to identify the presence or severity of sepsis and to find out the type of infection i.e., fungal, viral or local. Evaluation of response for therapy, guidance in therapy, predict complications of sepsis, prognostication and organ dysfunction development etc. are the other uses of biomarkers. There are many biomarkers has been used for past years i.e., C-reactive protein and procalcitonin etc. but the procalcitonin worked as best marker for prognosis. Procalcitonin is type of substance that can be produced by many types of cells. The normal range of procalcitonin is 0 to 0.2micro liter but if it exceeds in patient from the normal range then it is considered as that patient is having infection. The result of this biomarker can be still challenged because they don't have sufficient sensitivity and specificity (Pierrakos & Vincent, 2010). The different clinical factors have identified for sepsis, but these factors are not independently associated. Mostly bacteria is considered as the main reason but other microorganism can also cause sepsis like fungi, virus, parasites etc. Infection mostly affects the intraabdominal and respiratory sites (O'Brien et al., 2007).

1.2 Pathophysiology of sepsis:

The typical host response to infection is a complicated process that locates and inhibits bacterial invasion while initiating tissue repair. It includes the development of anti-inflammatory and proinflammatory mediators as well as activate the phagocytic cells and control circulation. Sepsis occurs when host response to an infection become widespread and affects the tissues which are far from the infection site. The response to infection starts when macrophages (which are innate immune cells) bind to microbial components. It occurs by including several steps. There are some receptors present on the surface of immune cells known as pattern recognition receptors (PRR) bind to molecular motifs of microorganism i.e., pathogen associated molecular patterns. They are recognized by toll like receptors, retinoic acid inducible gene I (RIG-I) like helicase and leucine rich repeat proteins, named as nucleotide-oligomerization domain (NOD). For example,

lipopolysaccharide from Gram negative bacteria bind to CD14 complex which is lipopolysaccharide binding protein on host immune cells. PRRs can also be known as danger associated molecular patterns (DAMPs) which released during inflammation. DAMPs are mitochondrial structure acquire specific functions when released into extracellular environment. ATP metabolic molecules, heat shock, mitochondrial DNA are examples of DAMPs. When extracellular signals bind to microbial components then immune system starts to trigger. Some other cell structures may also release during infection that may affect host response. Microparticles emitted by circulating and vascular cells also contribute to the negative effects of sepsis induced intravascular inflammation. While formation of NET is an important strategy for immobilizing and killing invading microorganisms, NET release DNA, histones and bacterial proteins promotes thrombosis, inflammatory response etc. When receptors bind to components of microbes they show multiple effects, TLR activation initiates a signaling cascade by activating cytosolic nuclear factor-kb (NF-kb). When NF-kb is activated, it moves from the cytoplasm to the nucleus, binds to transcription sites, and activation of large number of genes e.g., chemokines (ICAM-1), interleukin-1(IL-1), proinflammatory cytokines (tumor) include in the host inflammatory response. PMNs (polymorphonuclear leukocytes) become activated and express fixed molecules, causing them to clump together and adhere to the vascular endothelium. There are some endothelium molecules that attract the leukocytes. PMNs pass through multiple steps to move towards the injury site. PMNs releases some mediators which cause inflammation to cardinal signals. This process is mixture of pro and anti-inflammatory mediators which is responsible of bacterial killing, phagocytosis of bacteria, phagocytosis of debris from the tissues which are injured, chemotaxis etc. If the pro and anti-inflammatory mediators balance each other than homeostasis can be restored and proposed result of tissue repair or healing. But the large quantity of cytokines in septic patient spread into bloodstream which cause development of sepsis. The cytokines include in occurrence of sepsis are interleukin-1(IL-1) and tumor necrosis factor alpha (TNFa) and in this condition the plasma level increases at earlier point then it eventually goes decrease at the level where it is undetectable. These cytokines are reason of activation of fibrinolysis, induction of proinflammatory cytokines, fever and hypotension. TNFa has vital role in sepsis i.e., circulation of TNFa with shock is higher in patients with sepsis than non-septic

patients, TNF α produced symptoms that are similar to septic shock. Binding of lipopolysaccharide with endotoxin is the reason of high level of TNF α in septic patients which transfer to CD14 and stimulates TNF α .

Pathophysiology

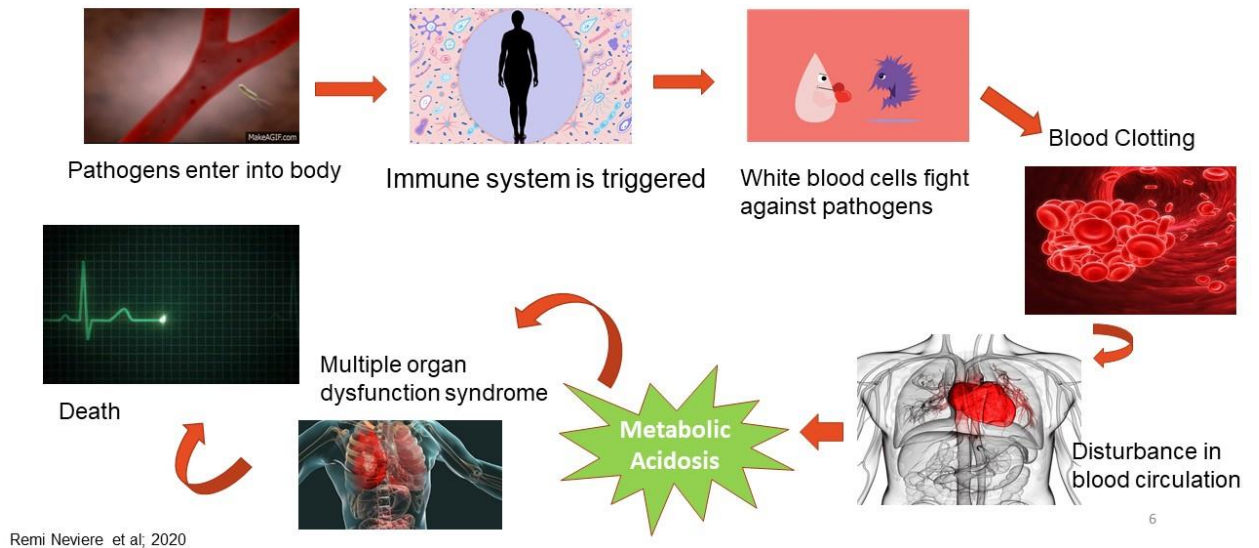


Figure 1: Pathophysiology of Sepsis

1.3 Multiple Organ Dysfunction Syndrome:

Multiple organ dysfunction syndrome (MODS) is a term that describes progressive organ disorder in a critically ill patients to the point where homeostasis can no longer be maintained without any intervention. It is at the high level of severity in both conditions infectious (septic shock, sepsis) and noninfectious. It can be classified as Primary MODS and Secondary MODS.

Primary MODS is a result of early stage of disease or infection (e.g. rhabdomyolysis cause renal failure). Secondary MODS is the result of host response (e.g., acute respiratory distress syndrome with pancreatitis). There are no criteria which is accepted at universe level for single organ disability in multiple organ dysfunction syndrome.

Progressive abnormalities in parameters of organ are frequently used to diagnose MODS and for the prediction of ICU mortality rate these parameters are also used in scoring systems (e.g., SOFA or LODS). The parameters are:

Respiratory – Partial pressure of arterial oxygen (PaO₂)/fraction of inspired oxygen (FiO₂) ratio

- Hematology – Platelet count
- Liver – Serum bilirubin
- Renal – Serum creatinine
- Brain – Glasgow coma score
- Cardiovascular – Hypotension and vasopressor requirement.

The larger the number of organ failures, the higher the death rate, and the largest risk of mortality associated with respiratory failure (Neviere et al., 2016).

1.4 Assessment of Sepsis

The assessment of organ dysfunction severity can be extracted by different scoring systems that exposed abnormalities on the basis of laboratory data and clinical reports (Singer et al., 2020). The scoring systems that are used for detection of sepsis on the basis of different biological events are SIRS, qSOFA, SOFA, NEWS etc. Scoring systems can enhance clinical suspicion of sepsis and prompt doctors to perform interventions that are time sensitive.

1.4.1 SIRS:(Qingqing Mao et al., 2018)

SIRS is clinical syndrome of dysregulated inflammation. It can be occurred in different conditions related or not related to infection. Not related conditions include pancreatitis, thromboembolism, autoimmune disorders etc. Many experts has presented that this criteria has been used in hospitals for many years but its ability to detect death is very poor in comparison of other scoring systems.

SIRS criteria are given below:

- a) Temperature $> 38^{\circ}\text{C}$ or $< 36^{\circ}\text{C}$.
- b) Heart rate $> 90/\text{min}$.
- c) Respiratory rate $> 20/\text{min}$ or $\text{PaCO}_2 < 32 \text{ mmHg}$ (4.3 kPa).
- d) White blood cell count $< 12,000/\text{mm}^3$ or $4000/\text{mm}^3$ or 10% immature bands.

1.4.2 qSOFA:

qSOFA is an updated version of SOFA. If score > 2 then it shows poor outcome due to sepsis. qSOFA prefer specificity because it is failed to achieve high sensitivity because it excludes important attributes i.e., temperature, heart rate etc. But qSOFA may be appropriate for screening at later stage.

qSOFA is easy to calculate as it only includes three parameters.

- Respiratory rate $\geq 22/\text{minute}$
- Altered mentation
- Systolic blood pressure $\leq 100 \text{ mmHg}$

1.4.3 NEWS

In comparison of all scoring systems NEWS is specific and having similarity like SIRS and showing best results without any requirement of laboratories for the detection of sepsis, severe sepsis and septic shock. The parameters included in detection of sepsis are:(com & 2008, n.d.)

- Respiration rate
- Oxygen saturation

- Systolic blood pressure
- Pulse rate
- Level of consciousness or new confusion
- Temperature

1.5 Machine Learning:

Currently, available screening methods for sepsis i.e. systemic inflammatory response syndrome (SIRS), modified early warning systems (MEWS), qSOFA etc. are not enough for clear identification of sepsis (Islam et al., n.d.) Many researchers are concentrated on machine learning approaches for the excellent outcome and high accuracy which is superior to the every disease severity scoring systems. Basically, machine learning aims to develop algorithm that can learn and create models for prediction and data analysis which give rapid outcomes (Chibani & Coudert, 2020)

This current work was designed to adopt a real time machine learning algorithms linear learner, Xgboost, multilayer perceptron neural networks, Lightgbm and random forest to detect sepsis at the time when patient admitted in ICU, based on Physionet data collected from two hospitals. In ICU, patients are admitted due to different reasons, the recognition of early sepsis with various disease states (e.g. inflammation) is quite challenging because every disease in ICU shows similar instances (e.g. dysregulated host response), clinical criteria (e.g. change in vitals) and symptoms (e.g. fever) (Moor et al., 2021). Machine learning models have ability to learn predictive patterns in data that helps to handle the complexity and wealth of digital patient data, which in turn give valid predictions about patient having sepsis. The predictive patterns can be exposed either through supervised or unsupervised learning. The algorithms that involve labeled training data (e.g., patients have sepsis or not) to predict outcomes for unforeseen data is presented as supervised learning. In contrast, the data which has no labels and determine (known and unknown) patterns in the data is included in unsupervised learning.

Over the last years, many research have used a range of computational models to deal with the difficulty in prediction of sepsis at its earlier stage. The large number of features are retrieved from available attributes to train different machine learning models and improve their performance. After verification of the proposed algorithms, through 5-fold cross validation method build the final ensemble model is applied on public challenge database and make evaluation of this model on the hidden test set (Yang et al., 2019.). The early detection of sepsis resulted in proper monitoring and management of the patient leading to significant reduction in mortality rate.

Chapter 2

Literature Review:

(Usman et al., 2021.) proposed comparison between SIRS, NEWS and qSOFA for the detection of septic shock and severe sepsis by collecting data of adults from emergence department. By calculating sensitivity, specificity and AUC curve it proposed that NEWS gave accurate and rapid results in detection of septic shock and severe sepsis while qSOFA showed poor sensitivity rate and invalid tool for sepsis screening. Systemic Inflammatory Response Syndrome is always targeted for its low utility and specificity score and qSOFA works well in non-ICU patients.

There is another study (Brink et al., 2019) on the comparison of scoring systems which included the data of suspected sepsis (described as the culture collection in ED) patients in emergence department. The predictive outcome is validated by discrimination AUC and it found that News showed best performance by giving the prediction of 10-30 days mortality. The limitation of this study was they have used the data of one tertiary care center, and they didn't give gold standard definition of infection.

(Qingqing Mao et al., 2018) But the detection of sepsis can be delayed by using these scoring systems so there was a need of rapid algorithms which could predict sepsis before these scoring systems and give prediction before onset of sepsis. In USA, annually 750000 patients in hospitals are diagnosed with sepsis and one third showing high mortality rate. Moreover, the average stay of sepsis patient in hospitals is more than the patients with other conditions due to which it shows high cost which is estimated at US \$23.3 billion in USA annually. Therefore, early prediction or detection of sepsis helped to control the longer length of stay of patients and mortality rate. For this purpose, this study has proposed Insight tool by using six vital signs directly excluded from electronic health records that included the patients over age of 18 years. This algorithm was validated on the data of three public hospitals and Stanford Medical center which gave best performance of model.

Furthermore, this algorithm was also trained on MIMIC III data (Multiparameter Intelligent Monitoring in Intensive Care). There were many missing values in the data which were imputed by carry forward method. After imputation, the data used to train Insight classifier and predictions tested on sepsis onset. Then this classifier was compared with other scoring systems and shown that Insight showing best AUROC curve as compared to SIRS, MEWS and qSOFA. But the limitation of this paper was this model was only trained for specific data and they didn't show their methodology that which vital signs they have been used for detection or prediction of sepsis.

(Nemati et al., 2018) proposed that sepsis is the disease which cause high mortality, morbidity rate and cost of ill patients in ICU. But there is no valid system exists for the prediction of sepsis onset. So, this study validated algorithm for prediction which is (AISE) Artificial Intelligence Sepsis Expert algorithm. It included EMR data and calculated 65 variables hourly and then implemented AISE algorithm which predict sepsis onset 4 to 12hrs before to clinical reports and presented those attributes which having great impact on prediction. Prediction of performance is inversely proportional to predictive lead time. AISE model gave the AUROC curve in range of 0.83 to 0.85.

(Islam et al., 2019.) Globally, sepsis is major health problem but there is no innovative tool for the detection of sepsis. Therefore, different machine learning techniques has been used for early prediction of sepsis by excluding different clinical variables from the data collected from different databases i.e., PubMed, Google Scholar, Scopus etc. which help the doctors in treatment on time and decrease the mortality rate and length of stay of patients in hospitals by giving better results than the existing scoring systems. It quantifies the working of model by proceeding meta-analysis and showed pooled area under receiving operating curve for predicting sepsis 3 to 4 hours before was 0.89, specificity 0.72 and sensitivity 0.81 while pooled area under receiving operative curve for MEWS, SOFA and SIRS was 0.50,0.78 and 0.70.

(Goh et al., 2021) Sepsis is blood poisoning disease that's detection and diagnosis is still challenging due to ambiguous symptoms and signs. This study developed SERA algorithm by using both structured data stored in EMR systems and unstructured data which include radiological images and clinical notes. In clinical notes mining, the researchers has used

natural language processing to extract medical information, clinical workflow and medical events that is stored in EMR data. In this way NLP AI algorithm has developed which combine with the NLP analysis of physicians that help to improve the accuracy to predict the risk factor of sepsis. This SERA algorithm further linked with other two algorithms which are diagnosis algorithm (which detects algorithm at time of consultation) and early prediction algorithm (which gives the prediction of sepsis in the next 4 to 48 hours). This SERA algorithm was tested on clinical notes which predict sepsis before 12 hours to the onset of sepsis and got sensitivity of 0.87 and specificity of 0.87 and AUC curve of 0.94. Then the results of algorithm were compared with physician's report and showed the potential of algorithm is up to 32% and reduce false positive rates up to 17%.

(Q Mao et al., 2018) validated a machine learning model named as Insight involved Xgboost package for the prediction and detection of sepsis and severe sepsis 4 hours before the onset from six vital signs by using the data of USA. The cross validation 10-fold method has been used for verification the performance of model and minimize the overfitting. The Insight algorithm in comparison of SIRS, MEWS and qSOFA showed better outcomes with AUROC score of 0.92%. But the limitation of this tool is it works only on specific data, so there is need to develop model that can run in every type of data so that, the model can be used in different hospitals of different countries.

(Hou et al., 2020) The better outcomes of survival and better treatment of sepsis can be done by early prediction using flexible machine learning algorithms for prediction. This study proposed the development of Xgboost algorithm to predict mortality of 30 days and comparison of trained model with existing traditional methods. The MIMIC III data was split into two categories survival and death.

(Calvert et al., 2016) This study has shown the retrospective analysis of adult patients (MIMIC II data) which didn't have sepsis at time of admission in ICU. Sepsis is a disease which is mostly caused by bacterial infection but can also be the reason of microbial endotoxin, viral and fungal infection. Sepsis is basically defined as SIRS with addition of suspected infection while severe sepsis linked with organ dysfunction and septic shock is associated with hypotension. Since 1991, sepsis detection method has been changed which included screening labs that are slow and inaccurate. Many studies have shown that early

detection of sepsis through Early Goal Directed Therapy can reduce the severe risk of sepsis and septic shock, but recent studies questioned on existing methods. Therefore, this study developed Insight tool as a early and better performance screening technology. In hospital settings many alarm indicators have been detected for severe sepsis and septic shock. The data of the adult patients included in this paper who didn't meet with SIRS criteria at time of admission, even didn't detect after four hours of stay. That is why Insight algorithm has used to predict 3 hours before. It presented AUROC curve of 0.92 at 3 hours before of SIRS episode. The performance of Insight algorithm was then compared with PCT procalcitonin which is biomarker used as laboratory test for sepsis. The AUROC of procalcitonin is 0.85 while Insight achieved specificity and sensitivity rate of 81% and 90% in comparison of PCT assay which was 63% and 80%. The best thing about Insight is it can combine multiple measurements and can find the correlation between them which would help in existing homeostatic condition.

(Xuze Zhao & Qu, 2021) Sepsis is most dominant cause of high morbidity and mortality in ICU patients. Therefore, reliable model for predicting the sepsis was required. So, the purpose of this study was to develop extreme Gradient Boosting based model Xgboost which gave better prediction than the other existing machine learning models. The data was collected from MIMIC III database of the patients having age between 18 to 89. Insight is an artificial algorithm which presents AUROC curve 0.79 in the prediction of sepsis before 4 hours to onset. Then another tool has developed for prediction of sepsis which was SVM support vector machine. This model gives the predictions by two ways left align or right align early prediction which achieved the AUC score 0.85. After it in 2020, Cristopher introduced the convolutional network for the sepsis prediction which gave positive rate 1.0 and false positive rate 0.0. But all these models are not practically used because many accurate models belong to black box model which couldn't give information about reasons that why model classifies the risk level of patients. Some drawbacks of traditional machine learning methods are imbalance change in range, adverse stability, low prediction power, etc. Therefore, the novel machine learning model has been designed. Many studies have revealed that Xgboost ensemble multiple weak models to make precise model. This model has selected on the basis of sensitivity, AUC, specificity, precision and error rate. The limitation of this paper was it has used limited dataset that model needs to

evaluate on different datasets. Moreover, it didn't notice the time factors on the predictive outcomes.

(Zabihi et al., 2019) Sepsis is a disease that is associated with skin, gut and liver infections. Early prediction of sepsis can reduce the associated mortality rate. The missingness in data has been controlled by using new different features. The major methods that used to achieve the goal are feature engineering and classification. Then ensemble technique Xgboost has used as predictive model which is officially ranked as third place in PhysioNet challenge 2019 with utility score of 0.339 on test dataset.

(Taylor et al., 2016) Predictive analytics in form of heuristics and scoring system has been limited for using in clinical decision rules. By the development of CDR, analytical methods proposed model by using small set of variables and rules which could be easily calculated. It takes many years to develop and lack of ability to update new information even its already available. But new machine learning models are capable of using large number of variables from electronic health records and make predictions on the basis of these variables. In this proposed study, machine learning approach was compared with existing CDR methods that have been used for the prediction and gave surety of better outcomes. The data split into 20-80 percent for training and validation. The model was developed by using data of electronic health records having 500 clinical variables. This model then compared with classification and regression trees, logistic regression model and other predictive model by using AUC area under the receiver operating characteristic curve. The main purpose of this model was needed to deploy for local predictions in hospitals.

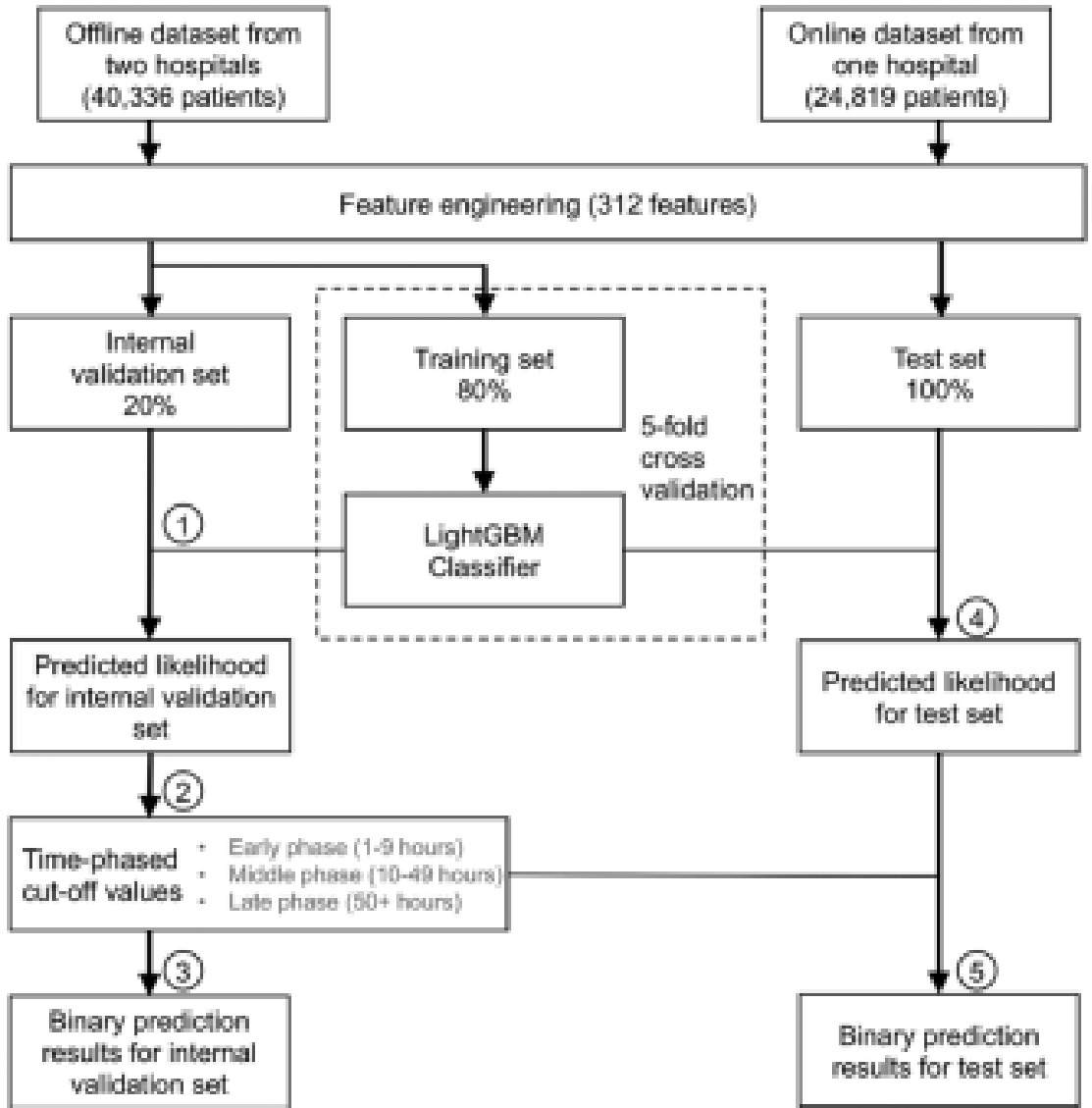
(Kong et al., 2020) The early detection of sepsis helps physician to make optimal treatment of ICU patients. The aim of this study was to propose the machine learning model to predict the risk of sepsis in ICU patients. For the development of model, MIMIC III data has been used which included 86 variables i.e. demographics and laboratory values. Different machine learning models random forest, logistic regression, least absolute shrinkage and selection operator, gradient boosted tree used for prediction. Then these models are compared with existing tools with Brier score, calibration plot, specificity and

AUC curve. SAPS II, APACHE II, III, IV (acute physiology and chronic health evaluation scores) are scoring systems used to assess the level of severity of sepsis in patients. These scoring systems are supposed to be best at time of development but with the passage of time and changes in population their performance became poor. The patients at the age of 18 to 90 years were included for the prediction. In this dataset mortality rate was 17.7%. The main purpose of this study was to predict sepsis during 24 hours after the admission in ICU. The ensemble methods which are based on decision trees are specific learning techniques use required parameters while logistic regression has ability to deal with high volume of data without distribution of patterns. While the gradient boosting tree and random forest ensemble weak decision trees and make a strong learner that perform better predictions. . Machine learning models have advantages to deal with high dimension data in which clinical variables have the impact of prediction in hospital mortality rate. The limitation of this paper is the data used in this paper is subset of MIMIC III data and has collected from single medical center. In the comparison of all machine learning models. Gradient boosting tree model is giving the best prediction as compared to random forest and showed better outcomes with high AUC score.

(X Zhao et al., 2021) Sepsis is basically out of control reaction of an infection which leads to high risk of death. In 2017, 48.9 million suffered from sepsis and people around 11 million died of sepsis. Two machine learning algorithms i.e., Xgboost and LightGBM are used to develop feature generation and mean processing methods that are used to predict sepsis 6 hours before of clinical reports. By combining window, medical and statistical features, feature engineering can be developed. PTT, platelets and white blood cells are considered as high-risk factors for prediction of sepsis which showed the inflammatory indicators. Vital signs having low proportion of missing values could easily measure but laboratory values having huge gap of intervals due to which there were large number of missing values. But to delete the missing values directly is not a good option because there is chance that useful information can be lost which is not valid for sepsis prediction so, this study has used Missforest method for imputation of missing values. Then 75 percent data used for training and 25 percent data used for verification in feature generation and mean processing method. LightGBM and Xgboost showed differ performance in mean processing method. LightGBM and Xgboost showed differ performance in mean

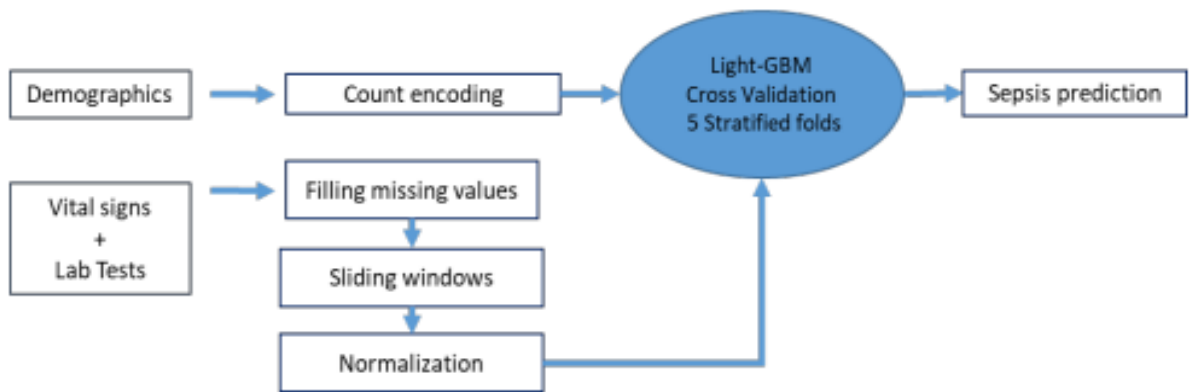
processing method. The recall rate of XGBOOST model is 0.55 with the different performance at 0-1 categories and its confusion matrix is in balanced way in test outcomes. While in feature generation method, both algorithms LightGBM and XGBOOST work well but in comparison LightGBM showed better recall and precision results in both categories of 0 and 1. LightGBM has fast speed of iteration as well as best predictive power because it works on leaf wise growth strategy which can easily deal with memory issue.

(Li et al., 2020) the real time prediction of sepsis has done in ICU by excluding dataset from PhysioNet challenge. It has also developed LightGBM model for the prediction by performing feature engineering. In every in every hour of stay in ICU. It randomly divided the data into 80% to 20%. To convert LightGBM into binary classification this study proposed new time phase machine learning model that set three cutoff values with ICU length of stay. For the model evaluation, effect of every feature having impact on prediction is calculated by Shapley Addictive explanation value. The incidence of sepsis occurred by doing the partition of time into three phases. In first phase 1-9hrs, the occurring rate of sepsis is higher than the 2nd phase which is 10-49 hours while in third phase after 50hrs the incidence rate has arisen rapidly. SHAP method (van Doorn et al., 2021) used to explain these prediction made at every instance by the models LightGBM. In this way new model TASP has been proposed for the real time predictions which also help in decision making for doctors. It showed the importance of every feature while LightGBM gave the exact rules for making decisions for the prediction as it works as ensemble boosted tree. The limitation of TASP model is its generality and stability must be thoroughly assessed in prospective situations.

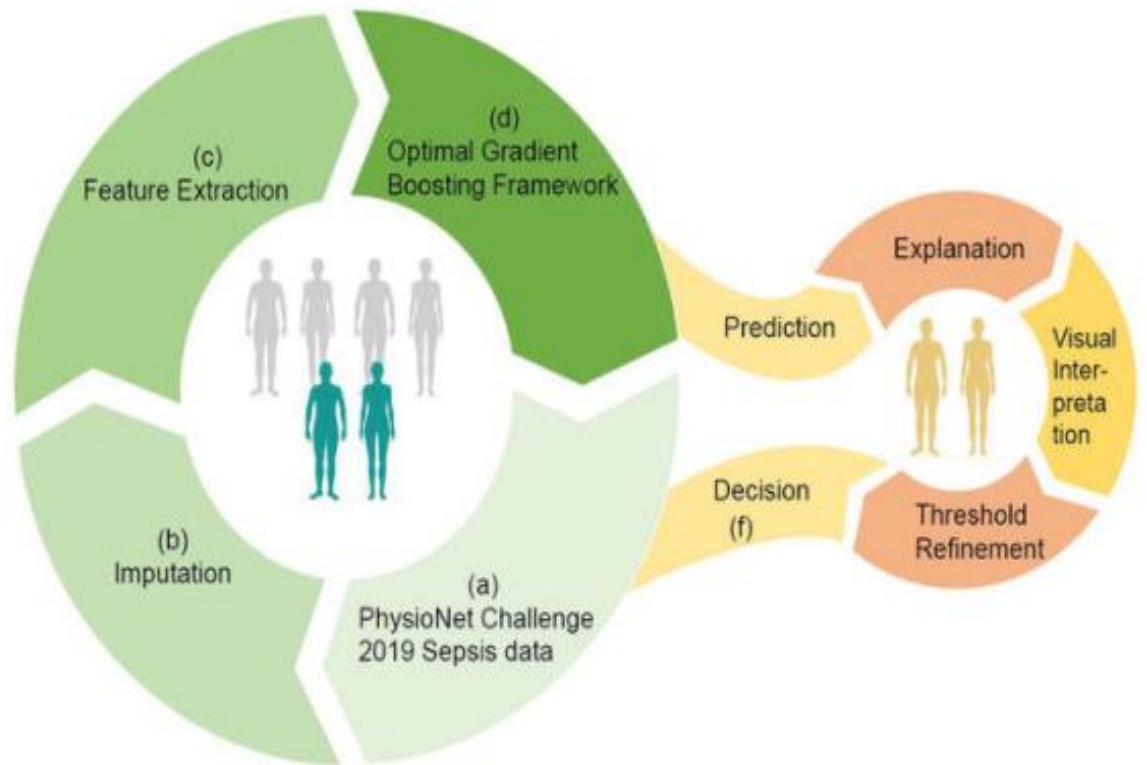


(Chami et al., 2019) By using artificial intelligence and machine learning, the recognition of sepsis can be faster, so the aim of this study to propose two methods, the first method is combination of neural networks and survival analysis and the other one is boosted tree method for the prediction 6 hours before of clinical reports. It included dataset which can be categorized into vital signs, laboratory and statistic values. As the data is collected by lab experiments so it's difficult to collect data hourly based. The deletion of missing values is not good idea especially of vital signs which are used for the prediction. Therefore, the

imputation of values has been done by using forward and backward values. Early prediction is basic application of Survival analysis that used in alarming events. So, SA can easily apply as statistical modelling technique to handle Time to Event problems. It is considered as Weibull Time to Event problem-RNN network (TTE-RNN) in which it is supposed that TTE follows Weibull distribution which further categorized into alpha and beta in this way they estimate distribution instead of variable. But this approach is not successful for prediction because there is still confusion that how this method is learning in data. LightGBM is then considered as best model for this approach.



(Nesaragi et al., 2021) The goal of this project is to create a machine learning model with clinical illustratable that can predict sepsis development before six hours and approve it with high-risk power of prediction for each time interval from ICU admission. The suggested approach allows for the study and applicable of clinical features for earlier prediction is explainable machine learning model for early prediction of sepsis xMLEPS. For each of the ten LightGBM models, a 10-fold cross-validation procedure is used high risk threshold for best prediction. Further these optimal models used related threshold values to improve predictive power using utility score for label prediction in every fold. The model was designed on publicly training data available, the complete framework is created using Bayesian optimization and trained with set of 85 features, giving an average balanced utility score of 0.4214 and 0.8591 area under the receiver operating characteristic curve.



Inter-relationships between clinical values have been shown to improve the ability of detection tasks. The physiological relations are obtained from the supplied variables after evaluating numerous research that establish the clinical importance of well-justified inter-relations among clinical symptoms. The imbalance data have been resolved by using LightGBM method with processing strategy.

Sl. no	Abbreviation	Description	Formula
1	SIndex	Shock Index (SIndex) is the proportion of heart rate (HR) being divided by systolic blood pressure (SBP), normalized by age.	$(HR/SBP) * Age$
2	DBPSIndex	Diastolic Shock Index is the proportion HR being divided by systolic blood pressure (DBP), normalized by age.	$(HR/DBP) * Age$
3	MAPSIndex	It is defined as the proportion of HR being divided by Mean Arterial Pressure (MAP), normalized by age.	$(HR/MAP) * Age$
4	BUNCr	It is the ratio of Blood Urea Nitrogen(BUN) to Creatinine	$BUN/Creatinine$
5	BILTCr	It is the ratio of Direct Bilirubin (Bilirubin_total) to Creatinine	$Bilirubin_total/Creatinine$
6	SaO ₂ -FiO ₂	It is the ratio of oxygen saturation of arterial blood in percentage (SaO ₂) to the fraction of inspired oxygen (FiO ₂).	SaO_2/FiO_2
7	PaO ₂ -FiO ₂	It is defined as proportion of the partial pressure of oxygen PaO ₂ divided by the fraction of inspired oxygen (FiO ₂).	PaO_2/FiO_2
8	Pla_Age	It is the ratio of platelets to age	$Platelets/Age$
9	PP	Pulse Pressure (PP) is the difference between SBP and DBP	$SBP-DBP$
10	CO	Cardiac Output is the product of pulse pressure (PP) and HR.	$PP * HR$

Three well-tuned baseline studies are undertaken as part of comparison analysis: The first one is, in 10-fold cross validation, the suggested technique is evaluated using a feature set of 85 characteristics without the use of optimal threshold refinement. In other methods, the 40 variables are directly trained in LightGBM model with or without check the threshold in cross validation technique. Then presented method xMLEPS used these three studies. The third study showed extreme results without optimal set of features and threshold. This study assures that data-driven automated ML models i.e., xMLEPS have the ability to alter the pattern from traditional detection to automated early prediction that prevents organ system failure due to sepsis.

(Adegbite et al., n.d.) examined performance of Systematic Inflammatory Response, quick Sequential Organ Failure Assessment (qSOFA), Universal Vital Assessment (UVA) and Modified Early Warning Score (MEWS) scores for prediction and diagnosis of death rate

with infection in underdeveloped and low-income countries. qSOFA is used as screening tool at very high-risk condition but with very poor outcome. While SIRS as suggested that not be used in severe sepsis because of its low sensitivity and specificity rate in finding patients with severe infection. SOFA cannot be applied outside the ICU because it requires laboratory values. All these tools mostly used in high income countries because low- and middle-income countries having limited resources and mostly patients are not admitted in ICU even in severe condition of diseases.

Scores	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)
qSOFA vs SIRS			
qSOFA	0.72 (0.58-0.82)	0.67(0.55-0.79)	0.74(0.68-0.78)
SIRS	0.88(0.79- 0.93)	0.34(0.25- 0.44)	0.56(0.40-0.76)
qSOFA vs MEWS			
qSOFA	0.58(0.35-0.78)	0.78(0.62-0.88)	0.73(0.63-0.79)
MEWS	0.74(0.58-0.86)	0.55(0.35-0.74)	0.69(0.65-0.74)

(Hsu et al., 2020) compared different machine learning models i.e. SVM support vector machine, KNN, RandomForest, Xgboost etc. for the prediction of sepsis by introducing the

novel methods of imputation on the basis of medical expertise and signal processing. But the sensitivity rate of every model is very low, it needs to be high for the best performance in every manner.

Chapter 3

Methodology:

This research aims to predict sepsis at the time of patient's admission in ICU by applying machine learning algorithms and extracted out the best model for the prediction. There are five steps involved to achieve the goal.

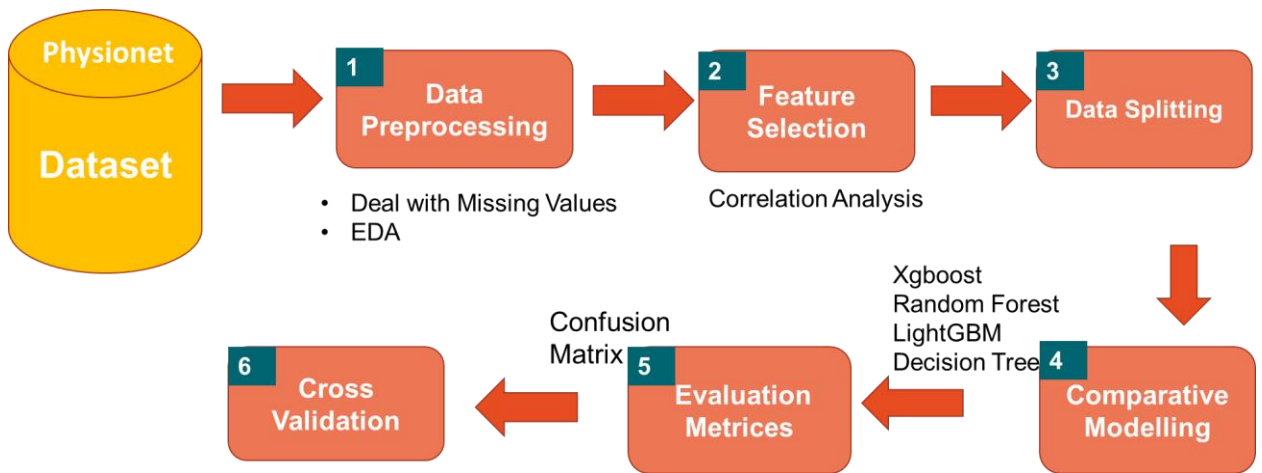


Figure 2: Steps of Methodology

3.1 Data Collection:

The data is extracted from Physionet challenge 2019 which consist of 40336 PSV files, collected from two different hospitals (Training set A which involved 20336 patients of hospital A and Training set B involved 2000 patients of hospital B). Each file indicates hourly recorded data of patients after admitting in ICU. The data includes 41 variables

which consists of 26 laboratory values (Measure of white blood counts, Bicarbonate, etc.), eight vital signs (temperature, heart rate, oxygen saturation, and systolic blood pressure etc.), six demographics (gender, age, ICULOS, etc.). The last variable represents sepsis label 0 and 1. 1 means the sepsis has identified in patient based on sepsis 3 criteria. The data is highly imbalance that only 2932 out of 40336 patients has sepsis. Additionally, there are many variables (26 out of 41) which have missing values more than 70 percent. For early sepsis prediction, the sepsis label has shifted forward for six hours in all data (meaning that the label is set to 1 for six hours before it is officially identified).

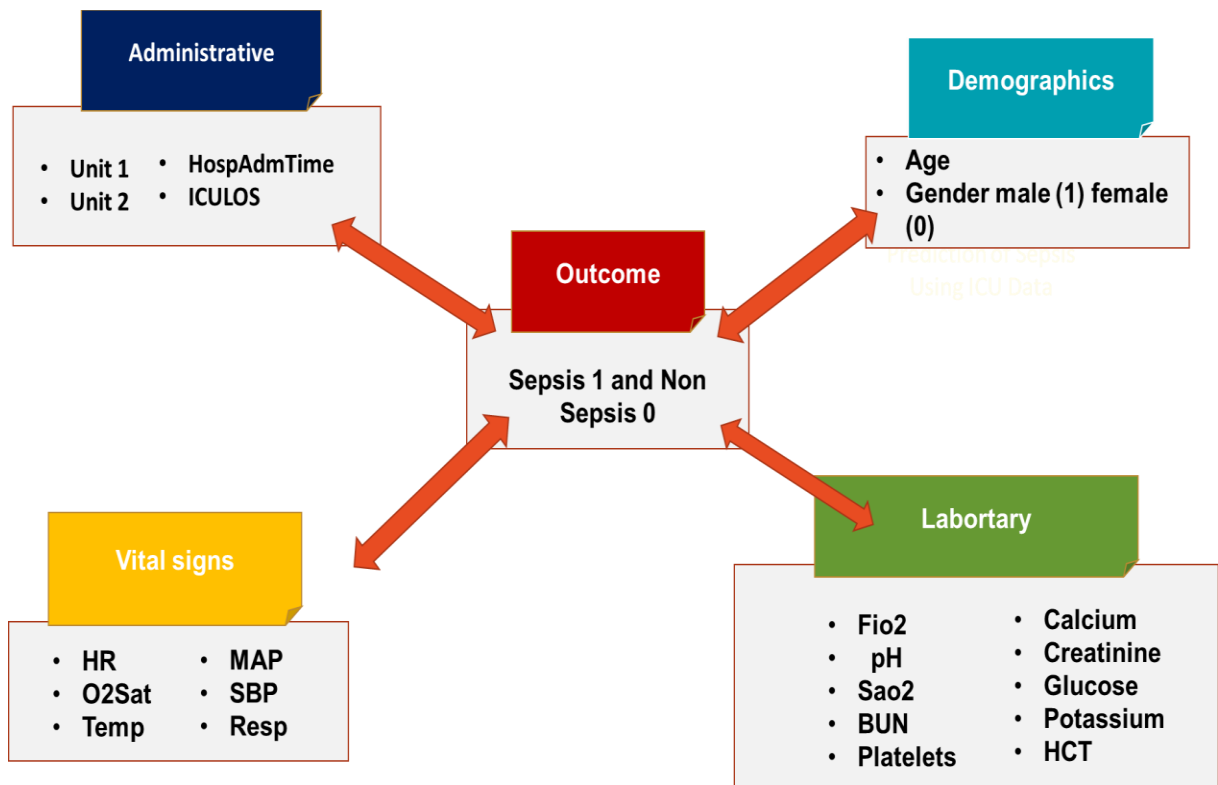


Figure 3: Data Description

3.2 Tools Used:

There are many machine learning libraries i.e., scikit-learn, NumPy, pandas, matplotlib which are open source, and use for classification, clustering, regression and dimensionality reduction. Scikit-learn is one of the most popular libraries which is used for evaluation of

model and useful to extract important features. If the dataset is highly imbalance, then it is considered as quite challenging, so to deal with the imbalance dataset there is library of Imbalanced-learn which offers multiple resampling techniques i.e., SMOTE analysis.

3.3 Data Preprocessing:

It is the most important phase in data formatting and data normalization. The review of data should be carefully analyzed to avoid misleading results. Therefore, interpretation for accurate data should be done before model building. The process of data preprocessing deals with redundant and noisy data and its strategies involved imputation of missing values and feature extraction. The large number of missing values in the dataset was needed to be imputed for better prediction outcomes by using different methods. Missing values in the data having great impact on the working of classifier. The main method to normalize the data is Min Max scalar or Expectation Maximization algorithm used to estimate parameters in the presence of missing data and different methods can be used for imputation of missing values i.e.

- Mean Imputation
- Median Imputation
- Mode Imputation
- 0 Imputation
- Pre and Next Imputation
- Missforest Imputation

3.3.1 Case Deletion:

This method shows the deletion of those attributes which have percentage of missing values, or it can delete all the rows which shows NAN values in every feature, but the problem is by applying this method huge data will be deleted which left very small amount of data for the analysis that is not reasonable. It's a statistical approach and by default it is

present in many programs. The importance and relation of attributes and instances with the targeted variable must be considered before deleting data because some columns or attributes have large number of missing values,

but they cannot be deleted because they presented huge impact on analysis. Case deletion must be applied when the data lost randomly (Acuña & Rodriguez, 2004).

3.3.2 Mean Imputation:

The one technique that frequently used for imputation of missing values is mean method (Wu et al., 2019). The NAN values are replaced with mean of each column, but it has one drawback that data can be skewed. So, in this case mean imputation is not good idea to fill NAN values. And the other disadvantage is it cannot deal with covariance between attributes and is not good for large amount of data The others are variance underestimated, correlation between attributes is negatively biased, sample size overestimated, and distribution of new values are wrongly presented. The mean imputation can be done by using the command `df. fillna (df.mean())` .

3.3.3 Median Imputation:

As mean is affected by outliers so in this scenario median can be used. The median of each attribute is replaced with the missing values (Biessmann et al., 2018). If the data is skewed, then median imputation is good choice for missing area. Imputation of median can also be done by using numerical data `df. fillna(df.median())`. It is suitable for smaller datasets. It cannot be used for categorical features. It's not accurate because it shows similar data which cannot be used for analysis.

3.3.4 Pre and Next Imputation:

(Goeij et al., 2020) There are some special methods for imputation of ordered data. In the case of pre and next imputation, the missing value is filled by taking the average of previous and next value. This method works for both numerical and nominal data.

3.3.5 Mode Imputation:

(Aljuaid et al., 2016) In this case, missing values are replaced with the most frequent value in the column. This imputation can be done in both numerical and categorical data. But the drawback of this imputation is data will not clear due to the repetition of same number in missing place of every column doesn't show the reasonable results.

3.3.6 0 imputation:

The another statistical strategy is to replace NAN values with 0 number which fills up every missing instance in attribute with 0 but the cons of this imputation are it may cause biasness in the data.(Jang et al., 2020.)

3.3.7 Missforest Imputation:

(Stekhoven et al., 2012) ML algorithm used for imputation of missing values is Missforest because better imputation is the basic key for the better performance of model. Missforest follows random forest algorithm which handle all missing values according to its requirement. This algorithm imputes mean or mode in first two iterations and then from the third iteration it fits random forest on the observed part and predict missing part based on observed part. This iterative process continuous until it met reasonable outcomes. It can handle different kind of data i.e., continuous and categorical. This algorithm doesn't need

hyperparameter tuning because random forest. It predict the values on the basis of original data distribution and also useful to fix the imbalance data (X Zhao et al., 2020) The reason of multiple iterations is, from second iteration, random forest work on best quality data that itself has imputed predictively. Missforest is considered as best imputation method because the one thing is it is easy to use, and the other thing is its error rate is 50% less than other alternative imputed methods. The advantages of using this algorithm are:

- It doesn't require data splitting or standardization etc.
- It is robust for noisy data as it has built in feature selection.
- It is nonparametric. It doesn't make any assumptions about the relationship between features.
- It has excellent predictive power.
- It can work with high dimension data.

3.4 Feature Selection:

The mechanism of feature selection is used to filter out the most relatable features with the variable which are needed to predict. The model accuracy can be affected by using inappropriate features showing maximum outlier detection. This study has focused on six vital signs by having that idea that these vital signs are present in all ICU patients and can be used for sepsis prediction. The statistical and correlation analysis has been used to extract the features that were showing highly contribution for the predicting variable.

3.5 Correlation Analysis:

This type of analysis shows the relationship between variables that how much variables are correlated with each other. It measures the strength between binary variables and shows its

direction. The range of result of this analysis is from -1 to +1 which basically known as correlation coefficient. The positive sign shows that two variables are correlated in positive manner while negative sign shows the correlation in negative manner. While correlation coefficient zero indicates that there is no association between two variables. If the variables are normally distributed, then Pearson correlation method can be used otherwise Spearman Correlation method is used because it is nonparametric in nature and it is robust in detection of outliers as compared to Pearson correlation method. Correlation Analysis is kind of significance test and stop at the calculation of coefficient. The relation between two variables cannot be only judged on the basis of strength and direction but it must be assessed by checking their significance level by applying the test of significance.

3.6 Statistical Analysis of Data:

The variables are described as counts and percentages. This analysis shows the evaluation of selected variables based on Z test. Z test is basically showing the proportion of mean between two variables when variance is known, and data is very large. It selected the p value >0.05 and presents that is there any mean difference between disease and normal variable. If it shows any difference and reject null hypothesis then that variable is considered as statistically significant and shows the normal distribution (Dong Wang et al., 2021) (Shimabukuro et al., 2017).

After the statistical and correlation analysis six vital signs has confirmed for the further process which are heart rate, temperature, oxygen saturation, respiratory rate, mean arterial pressure and systolic blood pressure and diastolic pressure. These variables having great impact in the prediction sepsis and can be used for model building.

3.6.1 Gender Analysis:

This analysis is required to know that the numbers of male and female who have sepsis and who have different diseases in whole dataset, which is helpful to know that sepsis mostly

effects the female as compared to male. The difference in male and female shows different hormone response to an infection. The septic male and female have high estrogen level and shows the severity of illness in females than males. Females with septic shock have high anti-inflammatory mediators while males have high tendency to maintain the health status (Eachempati et al., 1999). So, by knowing the biological events it proved that females have severe effect towards illness.

3.6.2 Sepsis Label 0 and 1:

In datasets there are two classes of sepsis label 1 the patients who are having sepsis and sepsis label 0, the patients who have no sepsis and admitted in ICU due to different reasons. So, for the analysis there is need to find the number of counts which are septic, and which are non-septic.

3.6.3 Age Analysis:

The prevalence of sepsis is disproportionately higher in the elder patients and the age of a person is an independent predictor of death. The elder patients are non survivors of sepsis. Mainly the sepsis effects the patients under the age of 60-80 years.

3.7 Train/ Test Split:

The data is divided into train test and validation. The training data is used to learn the model and then for validation of model, the test data and validation data is used. The range of train/test/validation data is 60%/20%/20%.

3.8 Smote Analysis:

(Liu et al., 2020) Down sampling and up sampling are two typical methodologies for dealing with datasets that are unbalanced. A reduced number of typical instances are chosen when majority class data is down sampled. Down sampling is good since it reduces overfitting effects, but too much down sampling will result in a loss of important information and lower the classifier's performance. Up sampling is the process of creating synthetic samples from the minority class in order to increase the number of samples in the minority class to the point where the number of samples in minority class becomes equal to the samples of majority class. Synthetic sample generation, on the other hand, is challenging and might lead to overfitting if the created samples are too similar to the originals. SMOTE is regarded as an efficient up sampling algorithm for generating synthetic samples. It firstly determines feature vector and its closet neighbor, and then take difference between them. Then it adds the random number with the feature vector to generate a new point on the line segment. SMOTE applies the topological qualities of neighborhood points present in minority class, rather than producing copies of previous samples. As a result, the classifier which is trained on SMOTE's synthetic data is less overfit.

3.9 Machine learning Algorithms:

There are many traditional methods i.e., laboratory test, qsofa score, SIRS etc. to detect sepsis but delayed in detection due to unclear symptoms cause the high mortality rate and increase the cost of hospitals therefore, there was need to predict sepsis earlier than clinical reports. For that purpose, different machine learning algorithms can be used for early detection with the high sensitivity and specificity rate. For example, Xgboost, Random Forest and Linear learner, LightGBM etc.

Xgboost is one of the best algorithms for the classification problem and shows accurate performance. It shows iterative phenomena and combine all the results extracted from weak

decision trees and gives the best prediction. In every iteration it is focused on misclassified observations. It includes the gradient boosted trees and construct the model. XGBoost also has an advantage over other machine learning approaches in that it makes no assumptions about data distribution and instead employs individual decision trees, which means it may not be affected from multicollinearity. Another advantage of ensemble approaches like XGBoost is that it can evaluate importance of features automatically from a trained prediction model, resulting in a score for the value of each feature in model's boosted decision trees. The higher an attribute's relative relevance, the more it is used to make crucial judgments in decision trees (Burdick et al., 2020).

(Montomoli et al., 2021) Meanwhile, XGBoost may process missing data automatically by assigning a default direction to null values. There is very low risk of overfitting while using Xgboost. To achieve the best XGBoost model performance, evaluation of hyperparameters was required, which included number of estimators, maximum depth and learning rates. The original dataset was randomly partitioned into five subsets for this investigation. One-fold was utilized as a testing subset, while the other four-fold were used to tune the hyperparameters, with 25 percent used for calibration and the remaining 75 percent subjected to four-fold cross validation with grid search. The hyperparameters selected that have the greatest area under the receiver operator characteristic (Yao et al., 2020) (Zabihi et al., n.d.).

There are many classification techniques for developing model by using huge data. Random forest is supervised learning that can be used for regression and classification problems. It was selected as the modern machine learning-based model, and it may be viewed as an extension of existing tree-based classifiers and make prediction from every sample and choose solution by using method of voting. It is an ensemble technique that eliminates over-fitting by averaging the results, which make it superior to a single decision tree. To address a single prediction problem, ensemble learning entails combining numerous models. Using ensemble learning, many models are created that learn for independent prediction (Shenoy, 2020).

Random forest was chosen over other machine learning techniques (e.g., support vector machines) because it is like CART and has advantages when dealing with EHR data. Random forest is an ensemble-based strategy that constructs several decision trees (i.e., "forest") at the training data to offset the constraints of decision trees. Each tree is built from a randomly selected subset of the original training data. A random subset of the entire number of variables is evaluated at each splitting node. By adopting the mode of decision-making, it can reduce the problem of overfitting.

LightGBM is great classifier for prediction which works 6 times faster than Xgboost. It learned about those attributes which having great contribution in prediction (CHAMI et al., n.d.). LightGBM-based gradient boosting system provides a special sparse data processing strategy, which is critical in classification challenge with class imbalance. It depends on histogram-based algorithms which reduces consumption of memory and speed up the training step. It combines advance communication networking for parallel learning. That is why it is also known as parallel voting decision tree algorithm. In each iteration, divide the training data into multiple machines and perform a local voting decision to select the top-k attributes and a global voting decision to receive the top2k attributes (Dehua Wang et al., 2017).

Linear Learner algorithm is used for binary classification. It is having an option of normalization for preprocessing. By turning on the normalization, it moves towards the smallest sample of the data and find out mean value and standard deviation for every label and attribute. But for binary classification, only features can be normalized. There are many optimization algorithms are involved which can be used to take control for optimization processes and help to deal with hyperparameters. When many models are trained in parallel manner, then they are compared with validation set to check which model is optimal. The optimal model gave the best F1 score and accuracy on the validation set.

The other deep learning algorithm used for classification in advance level is multilayer perceptron neural network which is also known as feed forward neural network which involves input layer, hidden layer and output layer in which unlimited data can be used. It doesn't only include vital signs, but also demographics or laboratory values. This algorithm doesn't make any assumptions about distribution of data. The most attractive thing about

this technique is it can trained as numerical models on new data (Gardner & Dorling, 1998). It basically consists of nodes or neurons having weights. Each neuron in MLP is connected to multiple of its neighbors, with varied weights expressing the relative importance of the various neuron inputs on the other neurons (Heidari et al., 2016). The imbalanced number of neurons in hidden layer may cause the overfitting but there is no specific method to find number of neurons. It is only dependent on trial and error method (Orhan et al., 2011).

3.10 Cross Validation:

The statistical method that is used to evaluate the performance of machine learning models. The difficulty with residual evaluations is that they don't show how the learner will perform better for the prediction of unseen data. To avoid this problem, the complete data set should not be used while training a learner. Before the training begins, some of the data is eliminated. After training, the removed data can be used to assess the learned model's performance on "new" data. This is the core concept behind the cross-validation method, which encompasses a wide range of model evaluation techniques.

3.10.1 Hold Out Method:

This method is simplest type of cross validation. It includes two datasets, training set and test set. The function approximator solely uses the training set to fit a function. Then, for the test data, the function approximator estimate the new output values. As before, the errors it generates are added up to provide the mean absolute test set error, which is used to assess the model. This approach has the advantage of being usually preferred to the residual method and taking no longer to compute. Its evaluation, on the other hand, can have a wide range of results. The evaluation may be substantially influenced by which data points are included in the training set and which are included in the test set, and so the evaluation may differ significantly depending on how the division is carried out.

3.10.2 K fold cross validation:

One option to improve on the holdout method is to use K-fold cross validation. The holdout approach is done k times after the data set is separated into k subsets. One of the k subsets is used as the test set each time, while the remaining k-1 subsets are combined for a training set. The average error for all k trials is then calculated. The benefit of this strategy is that it doesn't matter how the data is separated. Every data point appears exactly once in a test set, and k-1 times in a training set. As k is increased, the variance of the resulting estimate decreases. The drawback of this method is that the training algorithm must be rerun k times from the beginning, which implies that making an evaluation takes k times as long.

3.10.3 Leave One Out Cross Validation:

It is K-fold cross validation taken towards logical extreme with leave-one-out cross validation, when K is equals to N which is the number of data points in dataset. That is, the function approximator is trained on all the data save one point N times before making a forecast for that point. The average error is calculated and used to evaluate the model, as before. The evaluation provided by the leave-one-out cross validation error (LOO-XVE) is good, but it appears to be highly costly to compute on the first pass. Locally weighted learners, on the other hand, can make LOO predictions just as easily as they do conventional predictions.

Chapter 4

Results:

Three datasets has been used Training Set A which involved patients 790215, Training Set B with 761995 number of patients and the third dataset has been made by adding both training sets A and B which is 1552210. The datasets included 42 variables which includes laboratory values, demographics and vital signs which has large number of missing values. The missing values are needed to be imputed by using different methods.

4.1 Percentage of Missing Values in Training Sets:

There are large number of missing values in dataset but the percentage of missing values in vital signs has shown below:

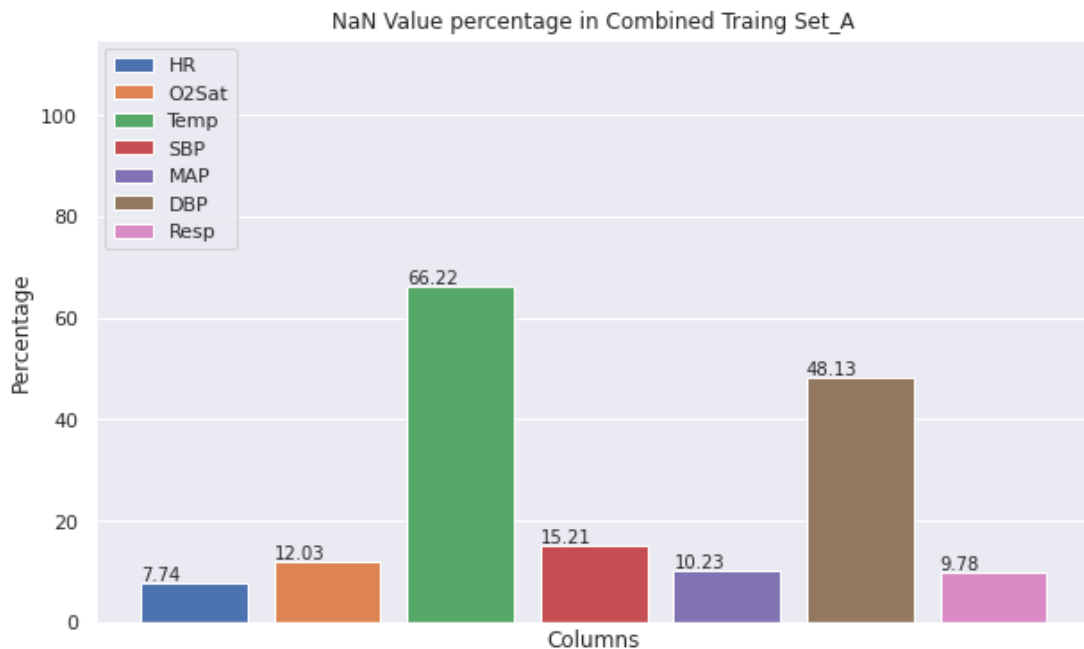


Figure 4: Percentage of Missing Values in Training Set A

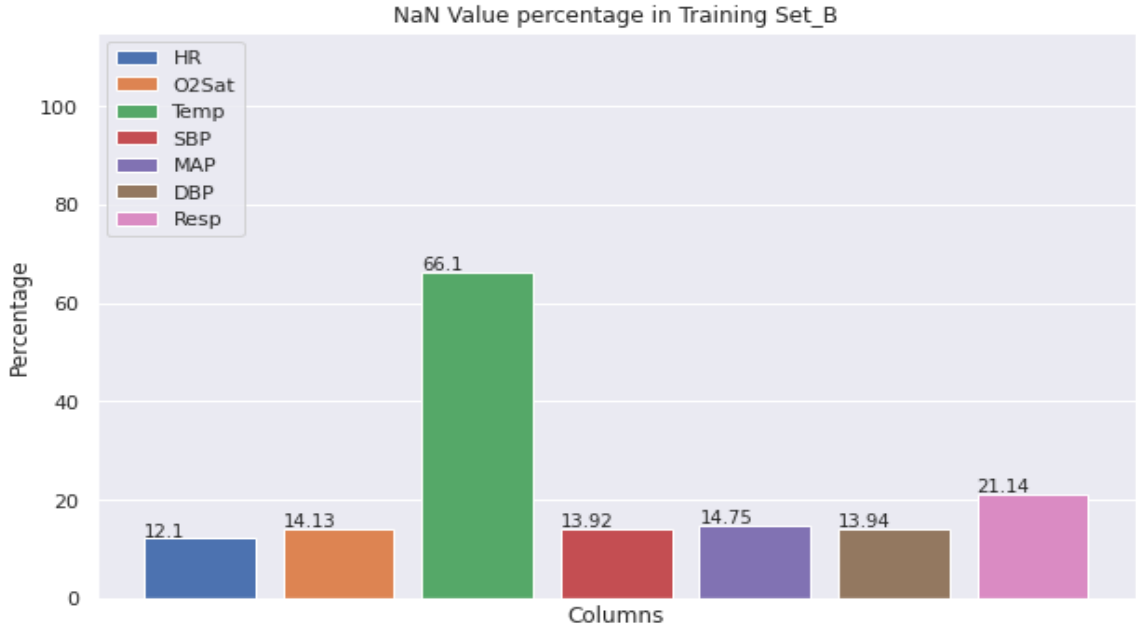


Figure 5: Percentage of Missing Values in Training Set B

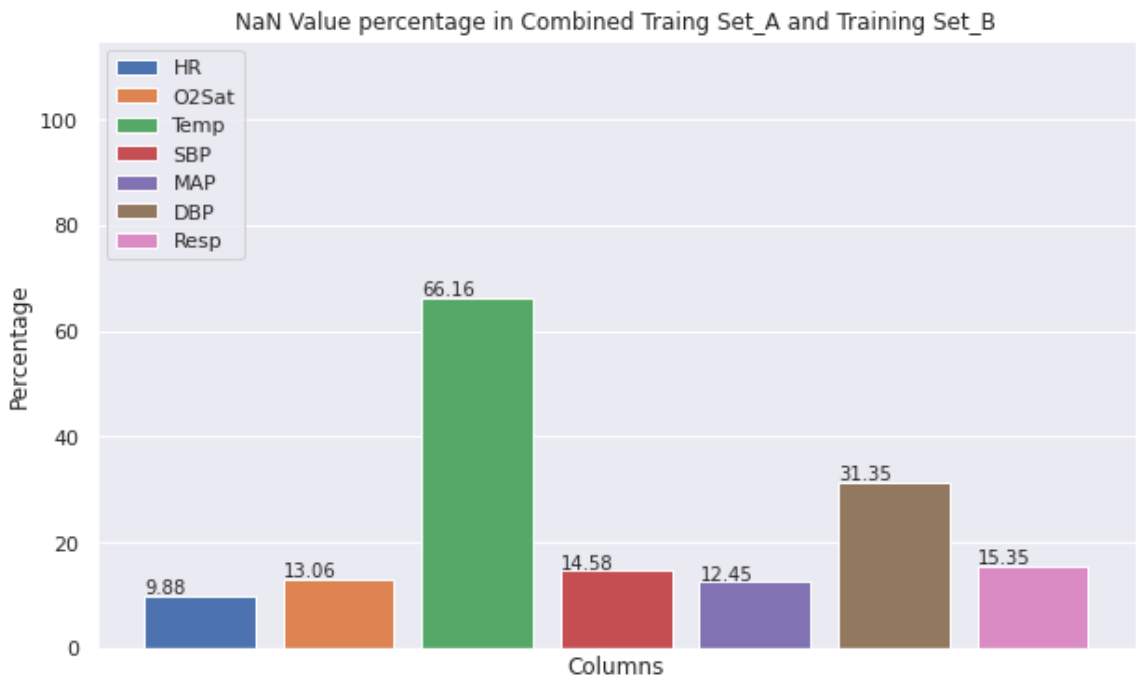
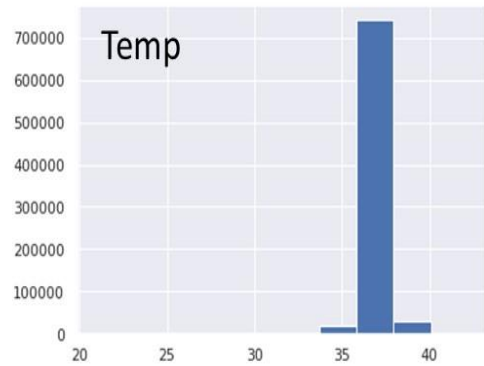
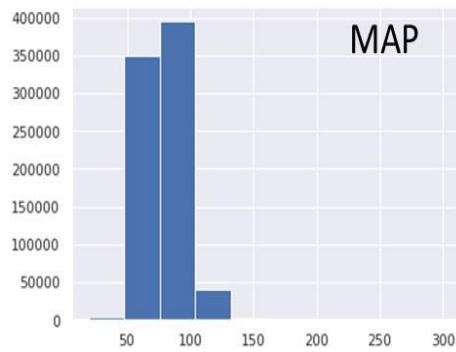
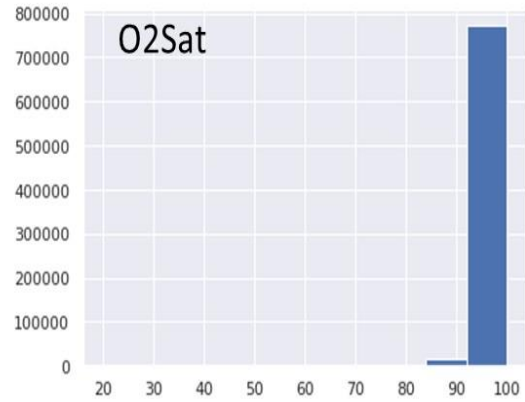
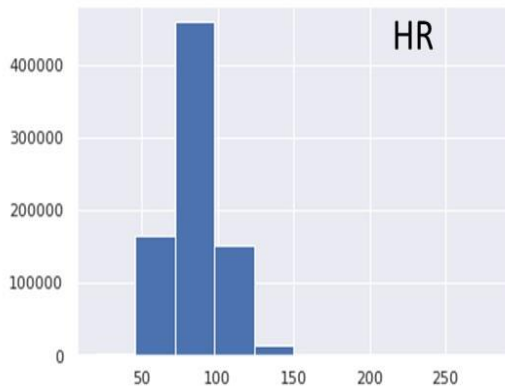


Figure 6: Percentage of Missing Values in Combined Dataset A and B

4.2 Histogram of Imputed Values in every column of Training Sets

After imputation of missing values through MissForest algorithm, every graph is showing maximum range of every column in every dataset.

Training Set A



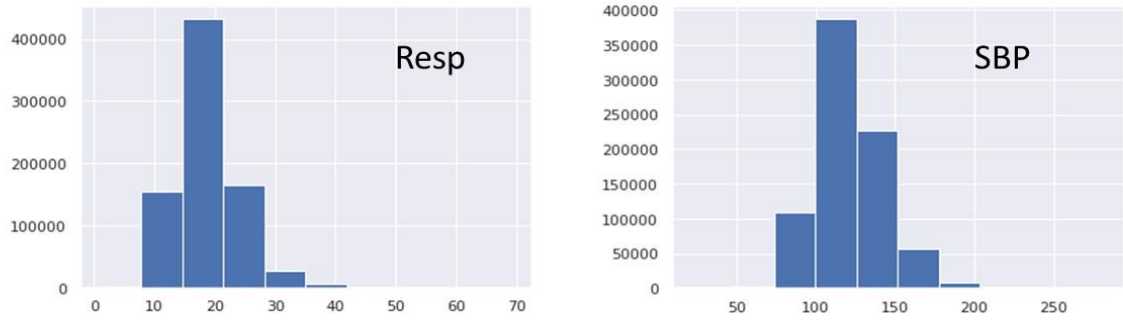


Figure 7: Histogram of Training Set A columns

Training Set B

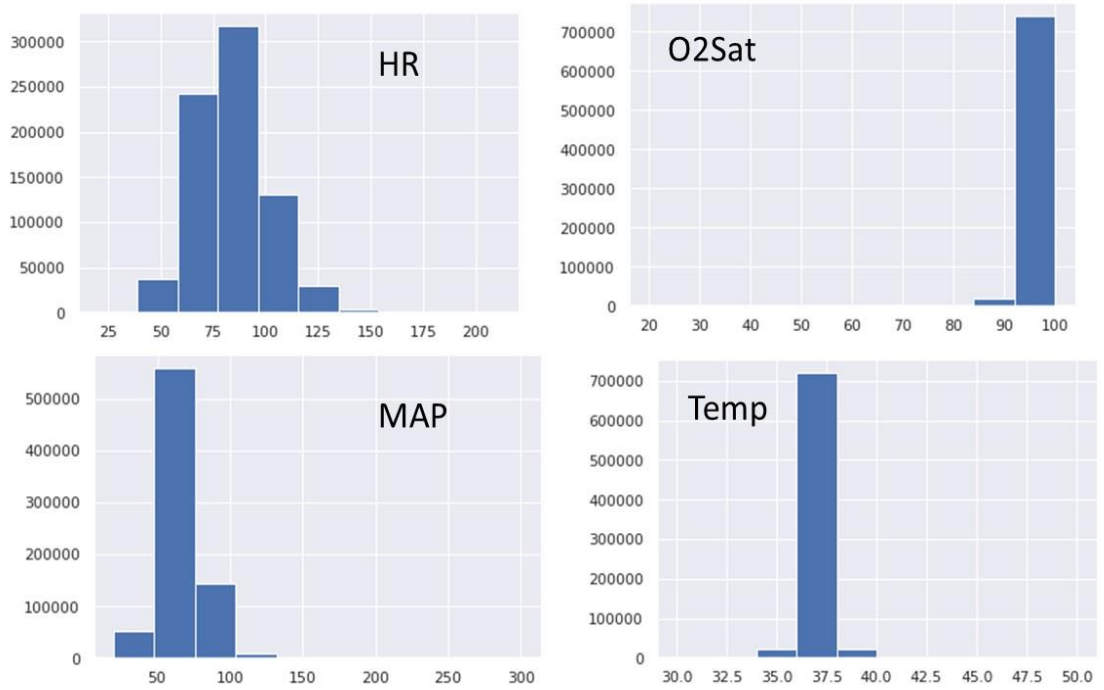




Figure 8: Histogram of Training Set B columns

Training Set AB:

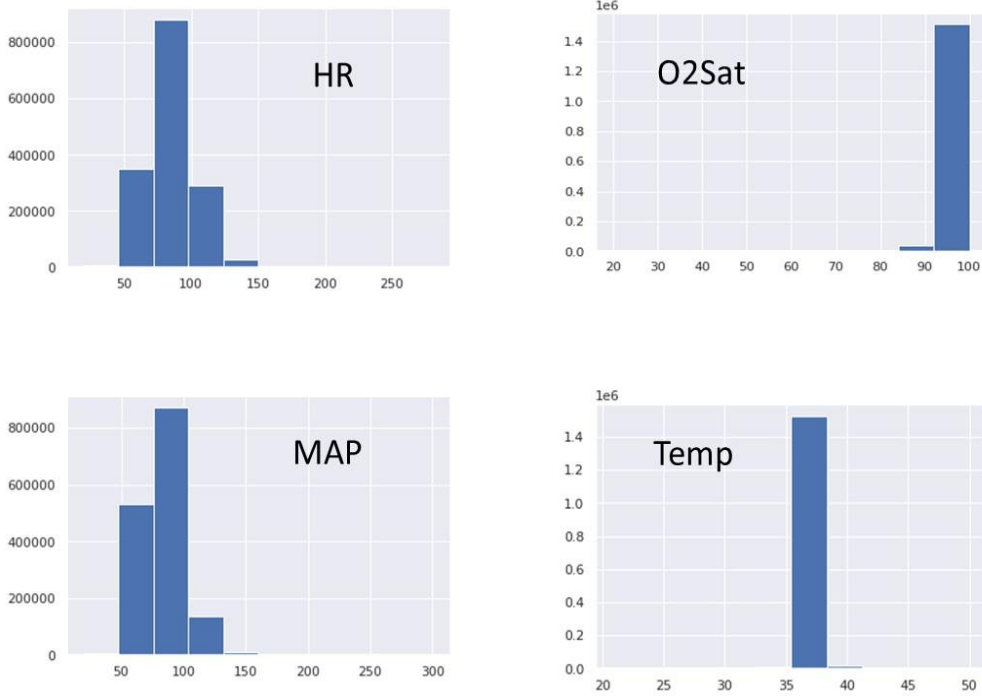




Figure 9: Histogram of Combined Training Set A and B

4.3 Correlation and Statistical Analysis:

4.3.1 Correlation Matrix:

Training Set A:

Spearman Correlation of Vitals in shifted_A_missforest_file

HR	1	-0.071	0.43	-0.028	0.12	0.22	0.25	-0.18	-0.03	0.02	0.03	0.0049
O2Sat	-0.071	1	-0.038	0.032	0.048	0.022	-0.2	-0.041	-0.048	-0.13	-0.0071	0.0046
Temp	0.43	-0.038	1	0.027	-0.0093	-0.0055	0.21	-0.16	0.082	0.18	0.04	-0.00017
SBP	-0.028	0.032	0.027	1	0.78	0.55	0.062	0.027	-0.0049	0.077	-0.0093	-0.0027
MAP	0.12	0.048	-0.0093	0.78	1	0.87	0.046	-0.19	0.033	0.029	-0.012	-0.0017
DBP	0.22	0.022	-0.0055	0.55	0.87	1	0.051	-0.37	0.08	-0.023	-0.0083	0.0059
Resp	0.25	-0.2	0.21	0.062	0.046	0.051	1	0.063	0.0087	0.12	0.032	-0.0056
Age	-0.18	-0.041	-0.16	0.027	-0.19	-0.37	0.063	1	-0.066	0.027	-0.002	-0.0096
Gender	-0.03	-0.048	0.082	-0.0049	0.033	0.08	0.0087	-0.066	1	0.00049	0.0057	-0.015
ICULOS	0.02	-0.13	0.18	0.077	0.029	-0.023	0.12	0.027	0.00049	1	0.058	0.01
SepsisLabel	0.03	-0.0071	0.04	-0.0093	-0.012	-0.0083	0.032	-0.002	0.0057	0.058	1	0.00085
Patient_ID	0.0049	0.0046	-0.00017	-0.0027	-0.0017	0.0059	-0.0056	-0.0096	-0.015	0.01	0.00085	1
	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	Age	Gender	ICULOS	SepsisLabel	Patient_ID

Figure 10: Correlation Matrix of Training Set A

Training Set B:

Spearman Correlation of Vitals in training_B_missforest_file

HR	1	-0.064	0.44	-0.02	0.052	0.11	0.23	-0.14	-0.015	0.066	0.029	-0.0098
O2Sat	-0.064	1	-0.064	0.023	0.031	0.018	-0.16	-0.068	-0.054	-0.096	-0.00053	-0.0045
Temp	0.44	-0.064	1	0.027	-0.11	-0.17	0.16	-0.068	0.046	0.21	0.04	0.01
SBP	-0.02	0.023	0.027	1	0.79	0.54	0.045	0.083	0.0035	0.059	-0.012	0.005
MAP	0.052	0.031	-0.11	0.79	1	0.87	0.06	-0.11	0.0048	0.048	-0.016	0.0039
DBP	0.11	0.018	-0.17	0.54	0.87	1	0.057	-0.24	0.063	0.026	-0.017	-0.00072
Resp	0.23	-0.16	0.16	0.045	0.06	0.057	1	0.05	0.0068	0.091	0.018	-0.0024
Age	-0.14	-0.068	-0.068	0.083	-0.11	-0.24	0.05	1	-0.017	0.014	-0.00046	-0.0016
Gender	-0.015	-0.054	0.046	0.0035	0.0048	0.063	0.0068	-0.017	1	0.0083	0.0038	0.019
ICULOS	0.066	-0.096	0.21	0.059	0.048	0.026	0.091	0.014	0.0083	1	0.047	0.009
SepsisLabel	0.029	-0.00053	0.04	-0.012	-0.016	-0.017	0.018	-0.00046	0.0038	0.047	1	0.00022
Patient_ID	-0.0098	-0.0045	0.01	0.005	0.0039	-0.00072	-0.0024	-0.0016	0.019	0.009	0.00022	1
	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	Age	Gender	ICULOS	SepsisLabel	Patient_ID

Figure 11: Correlation Matrix of Training Set B

Training Set AB:

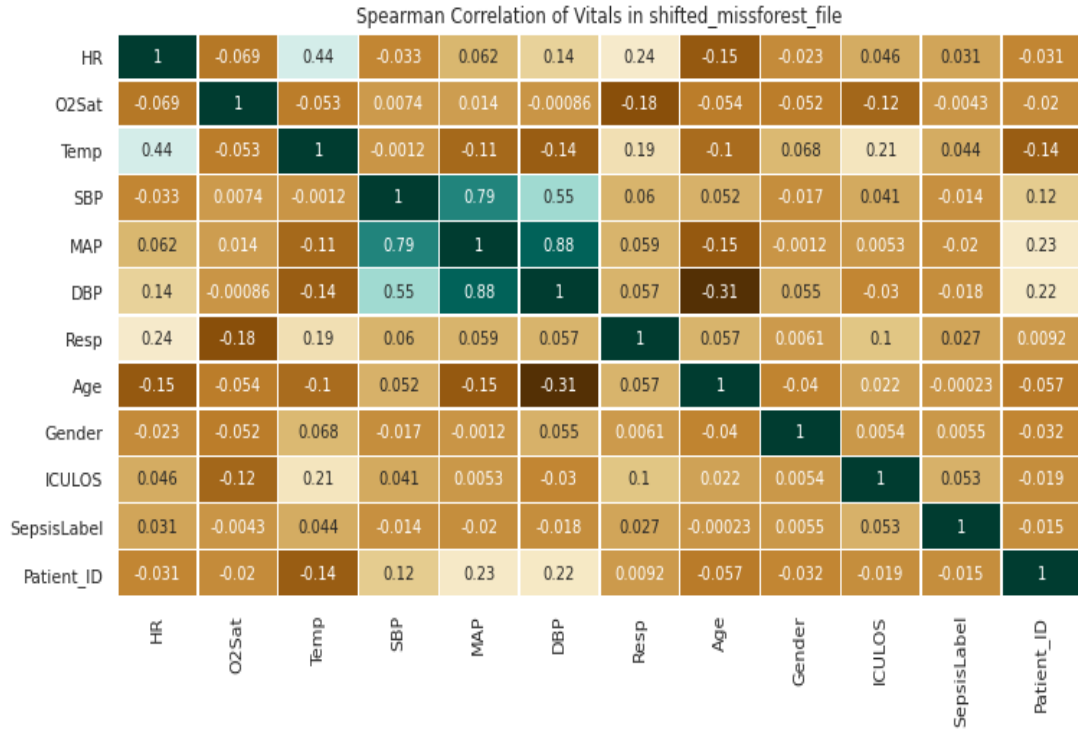


Figure 12: Correlation Matrix of Combine Training Set A and B

4.3.2 Gender Analysis:

This analysis is showing the number of males (0) and females (1) in the different datasets.

Training Set A

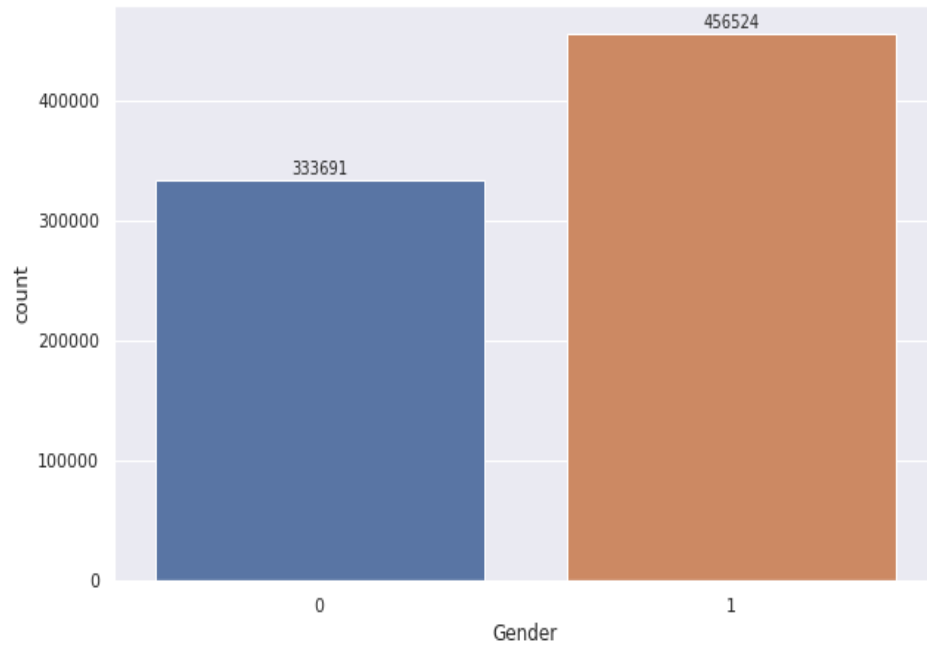


Figure 13: Gender Analysis of Training Set A

Training Set B:

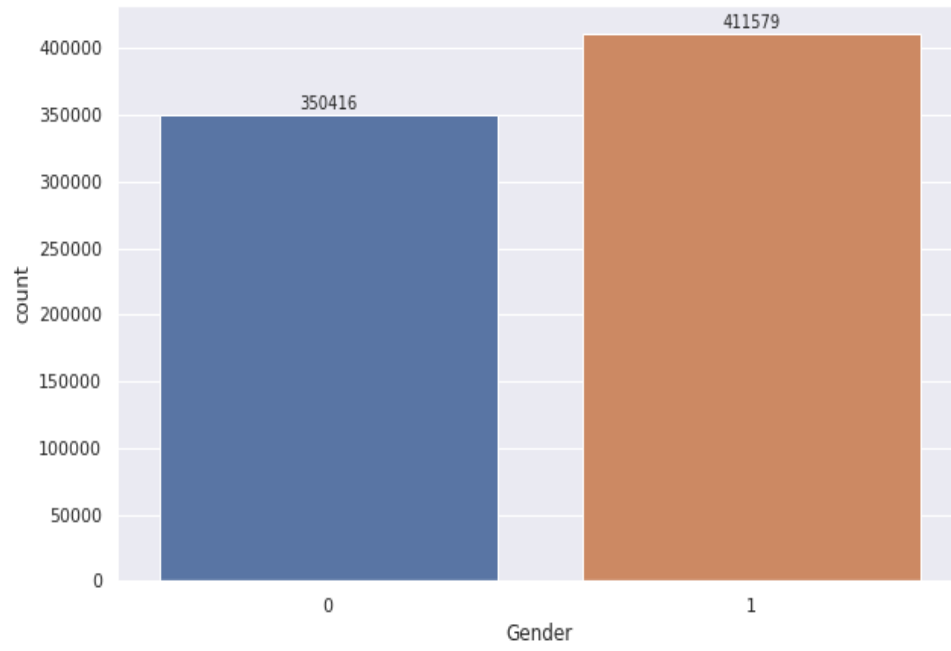


Figure 14: Gender Analysis of Training Set B

Training Set AB:

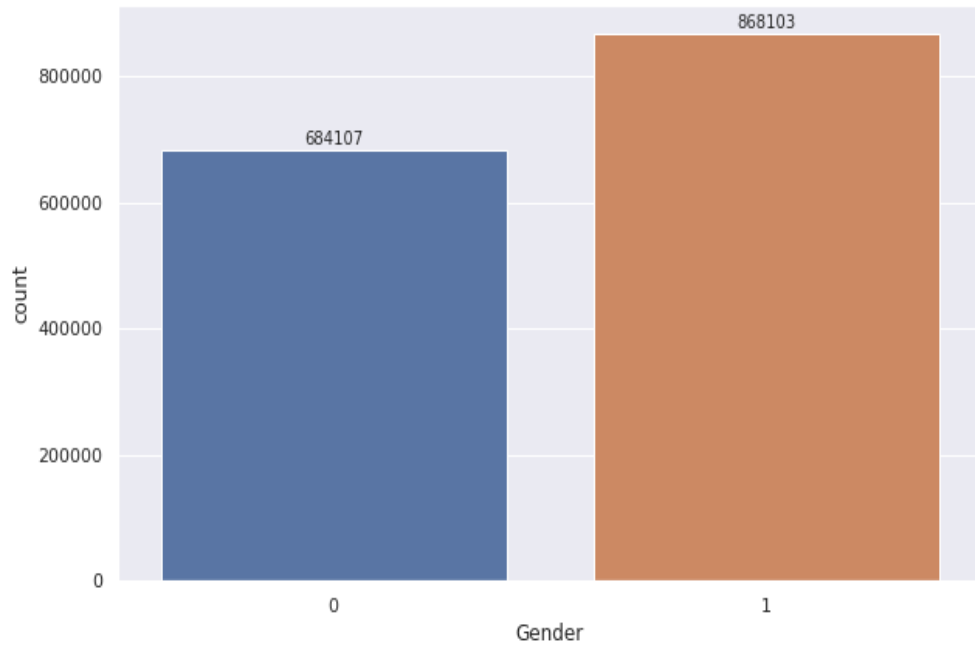


Figure 15: Correlation matrix of Combined Training Set A and B

4.3.3 Age Analysis:

Age analysis shows that at which range of age, patient having sepsis.

Training Set A:

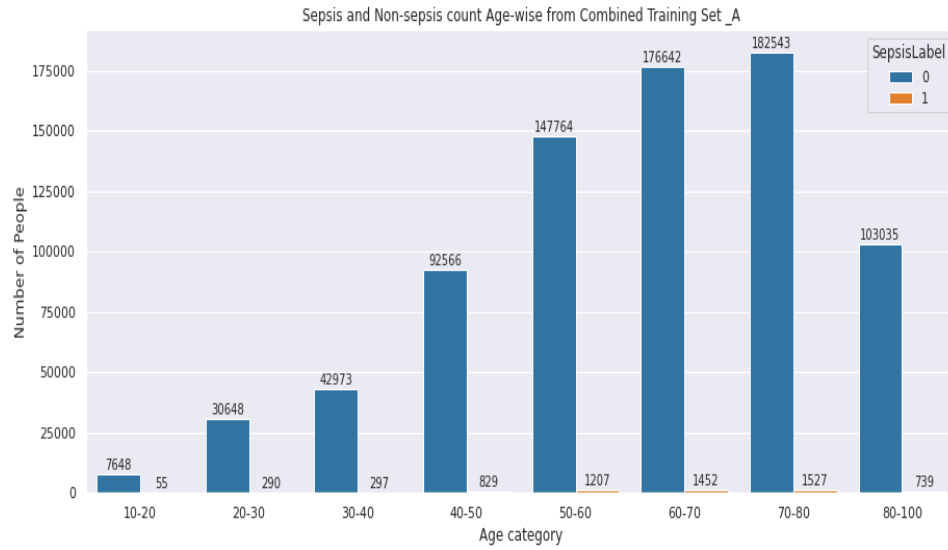


Figure 16: Age Analysis of Training Set A

Training Set B

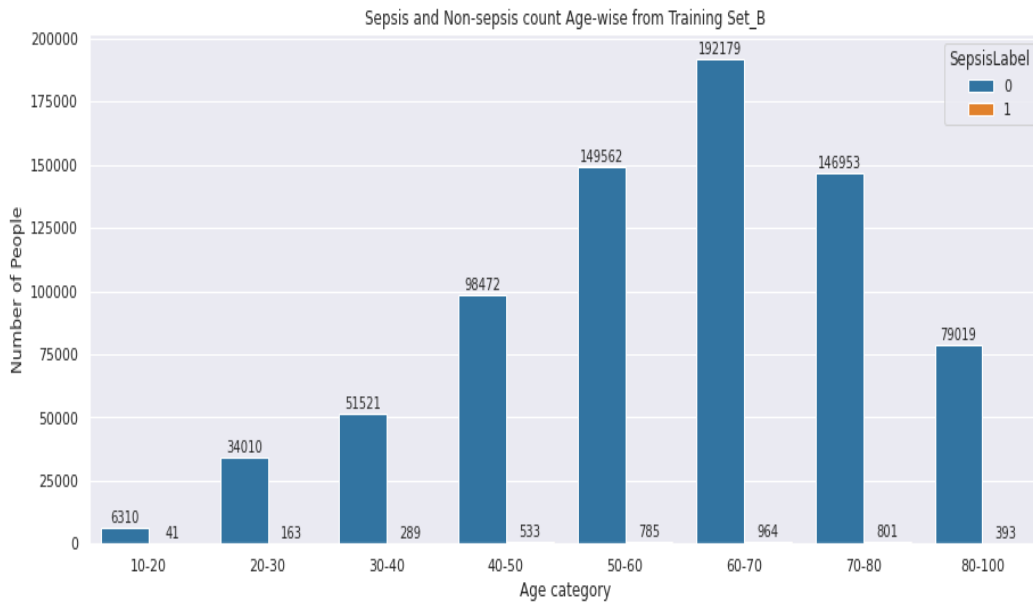
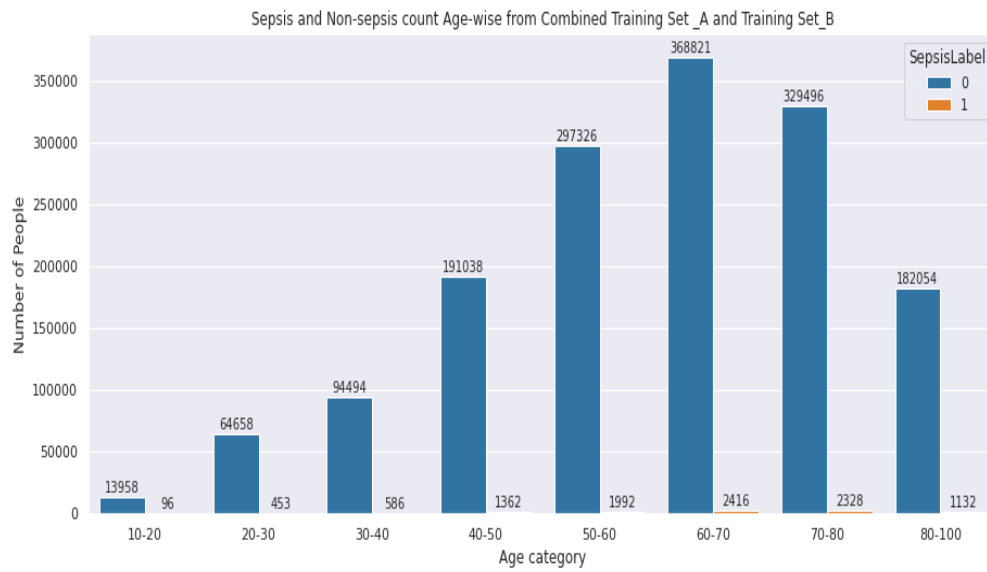


Figure 17: Gender Analysis of Training Set B

Training Set AB:**Figure 18: Gender Analysis of Combined Training Set A and B****4.3.4 Septic and Non Septic Patients:**

The number of septic and non-septic patients in datasets.

Training Set A:

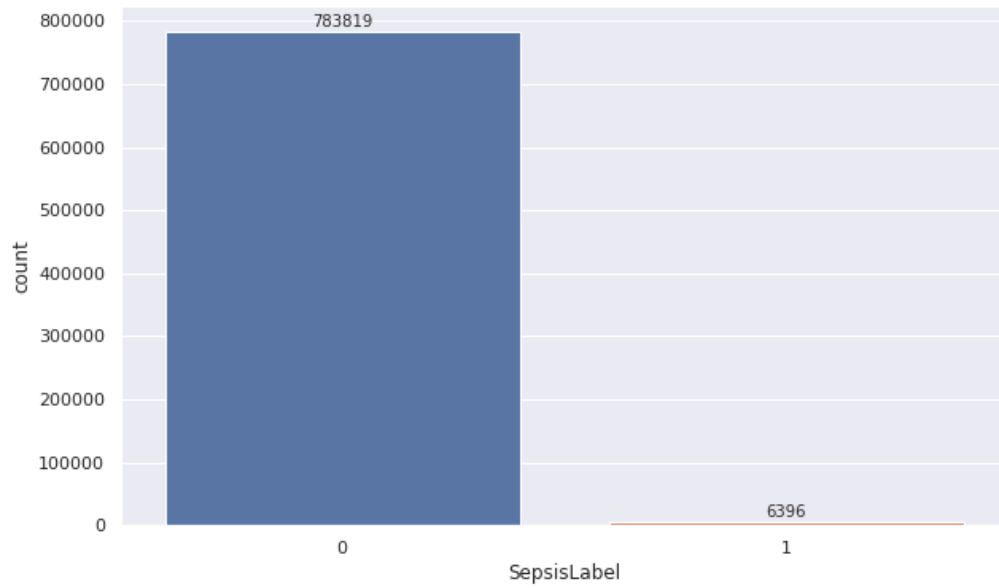


Figure 19: Number of Septic and Non-Septic patients in Training Set A

Training Set B:

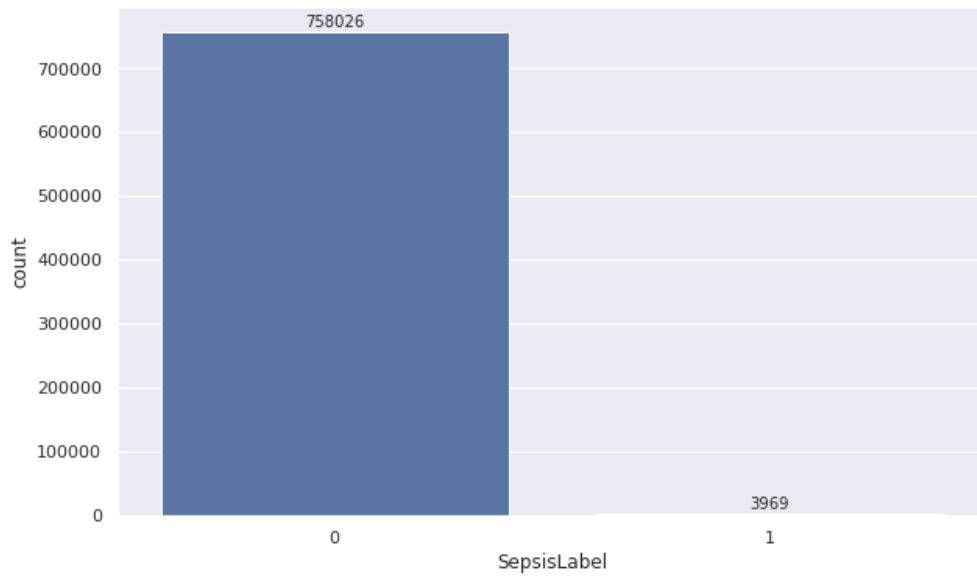


Figure 20: Number of Septic and Non-Septic patients in Training Set B

Training Set AB:

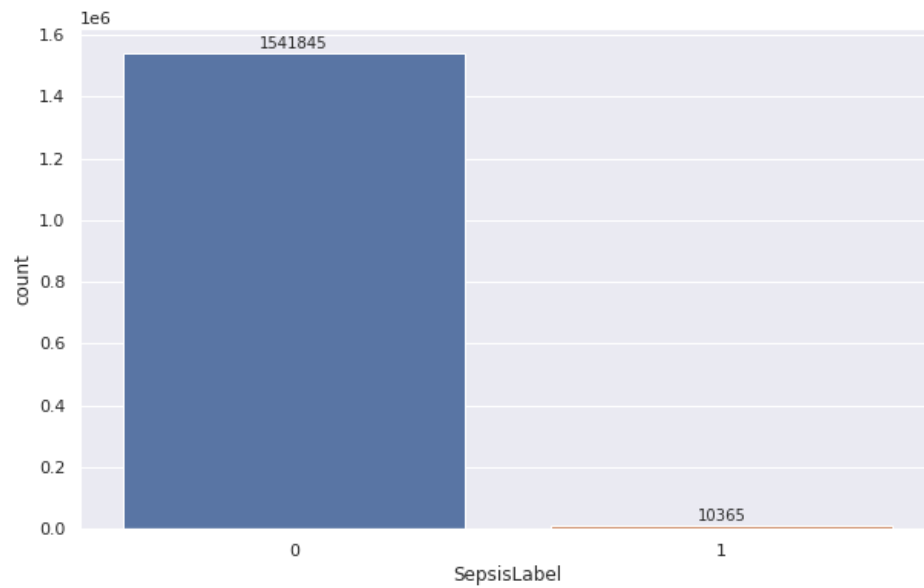


Figure 21: Number of Septic and Non-Septic Patients in Combined Dataset A and B

4.4 Confusion Matrix after Smote Analysis:

Confusion matrix shows the number of true positive ,true negative, false positive and false negative which means that how many patients are truly predict which are having sepsis and non-sepsis and how many patients are negatively predict having sepsis but in real they are normal.

Training Set A:

Xgboost

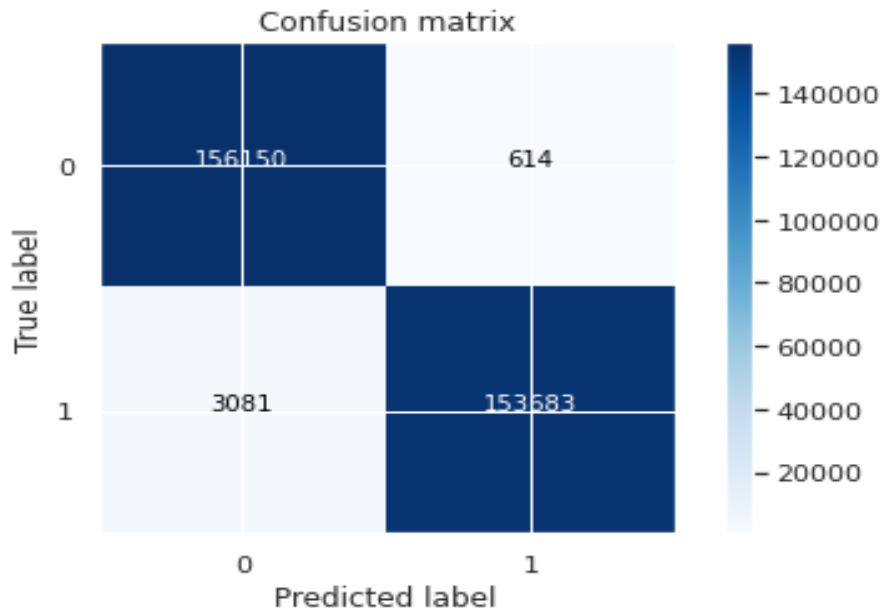


Figure 22: Confusion Matrix of Xgboost in Training Set A

LightGBM:

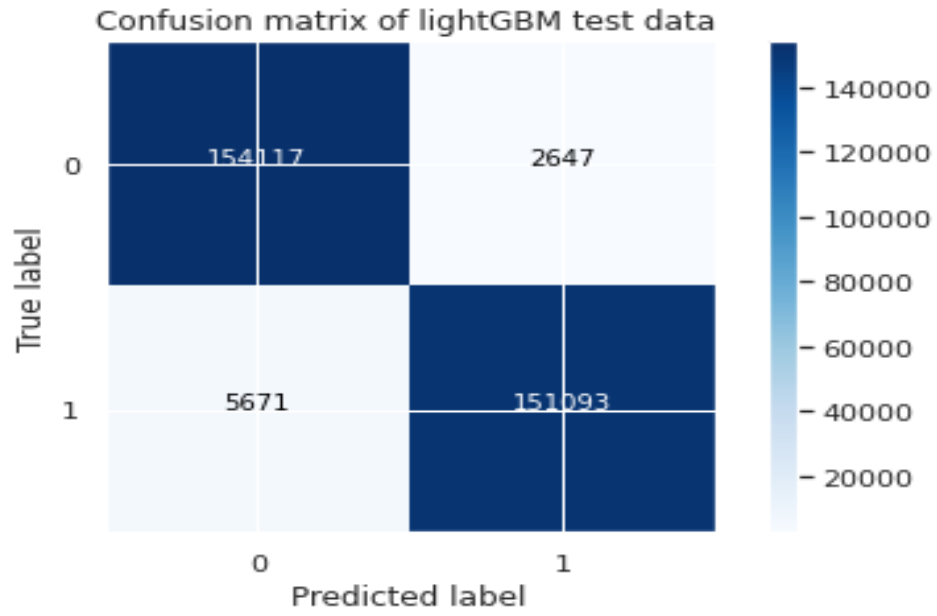


Figure 23: Confusion matrix of LightGBM in Training Set A

Random Forest:

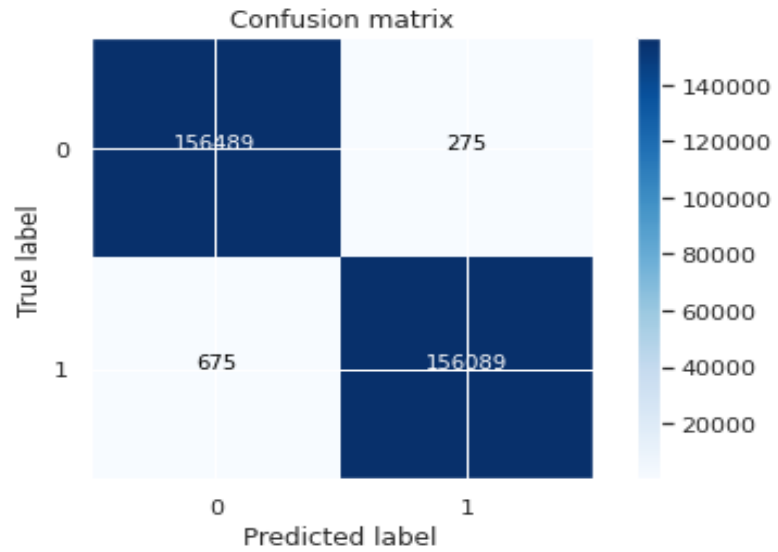
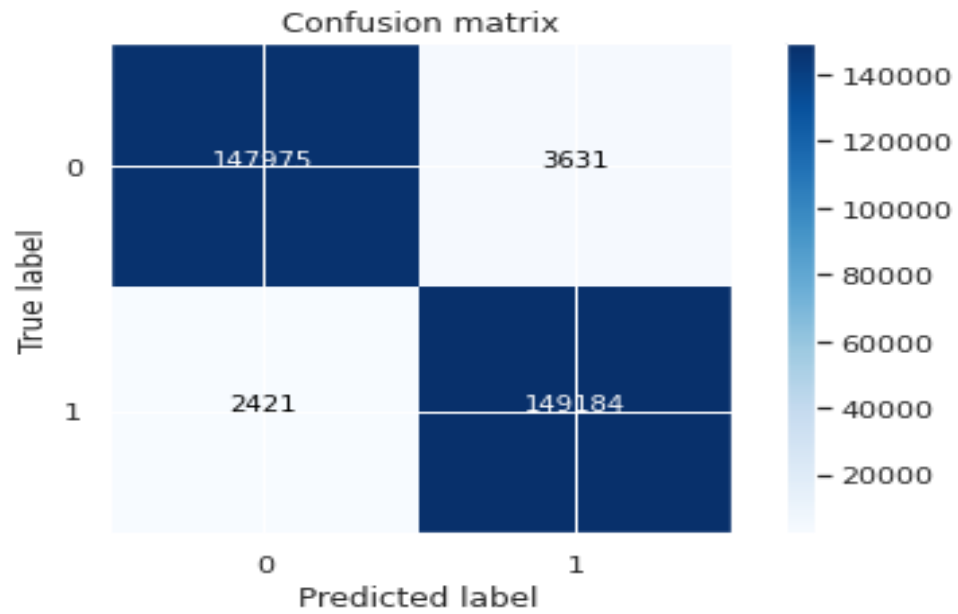
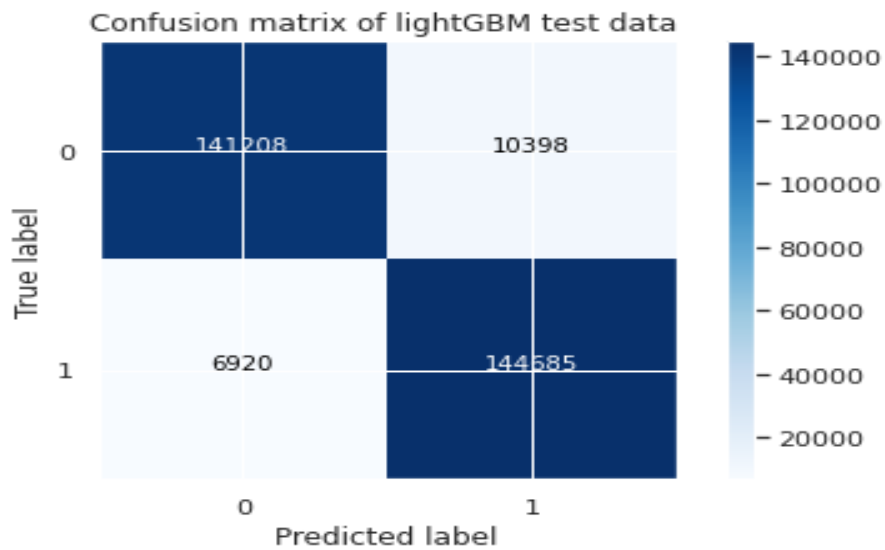


Figure 24: Confusion matrix of Random Forest in Training Set A

Training Set B:**Xgboost:****Figure 25: Confusion Matrix of Xgboost in Training Set B****LightGBM:****Figure 26: Confusion Matrix of LightGBM in Training Set B**

Random Forest:

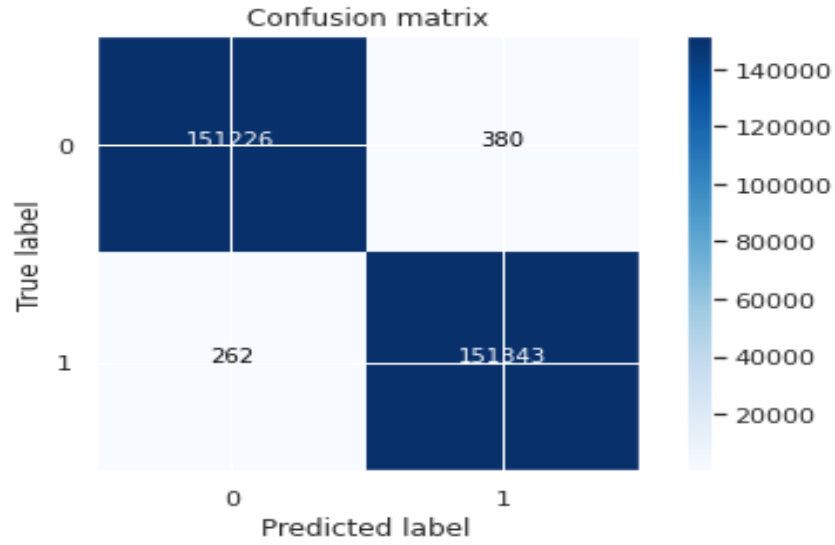


Figure 27: Confusion Matrix of Random Forest in Training Set B

Training Set AB:

Xgboost:

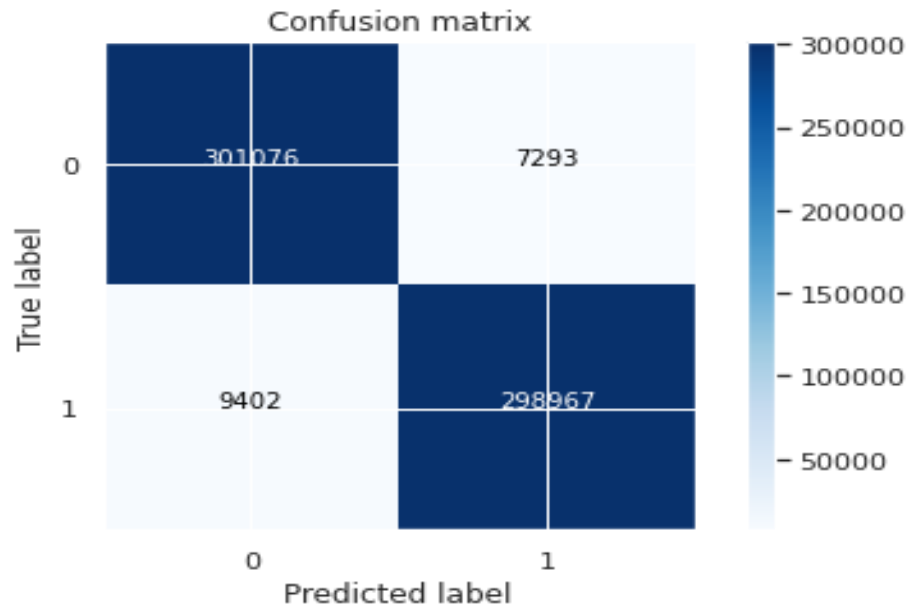


Figure 28: Confusion Matrix of Xgboost in Combined Training Set A and B

LightGBM:

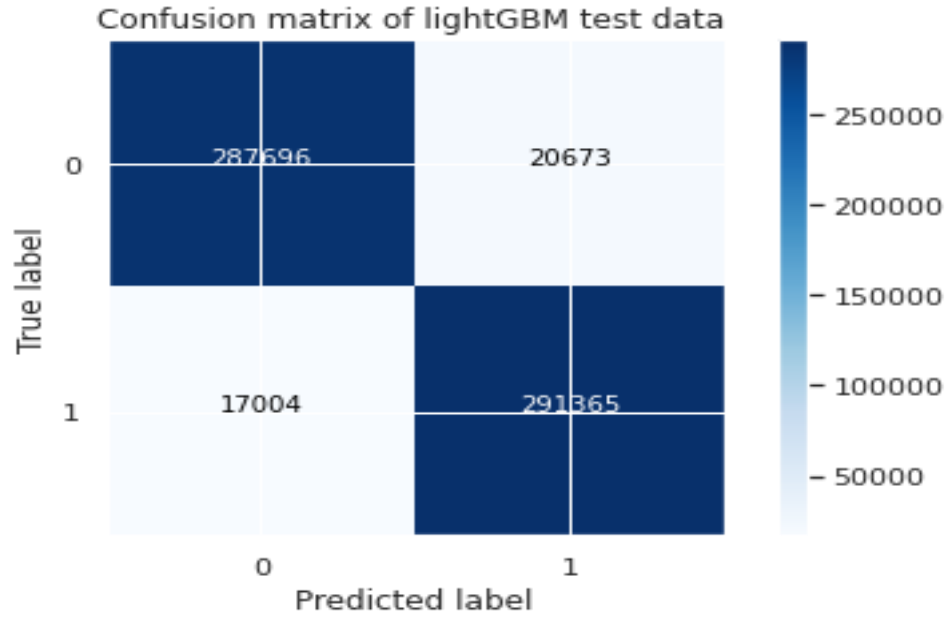


Figure 29: Confusion Matrix of LightGBM in Combined Training Set A and B

Random Forest:

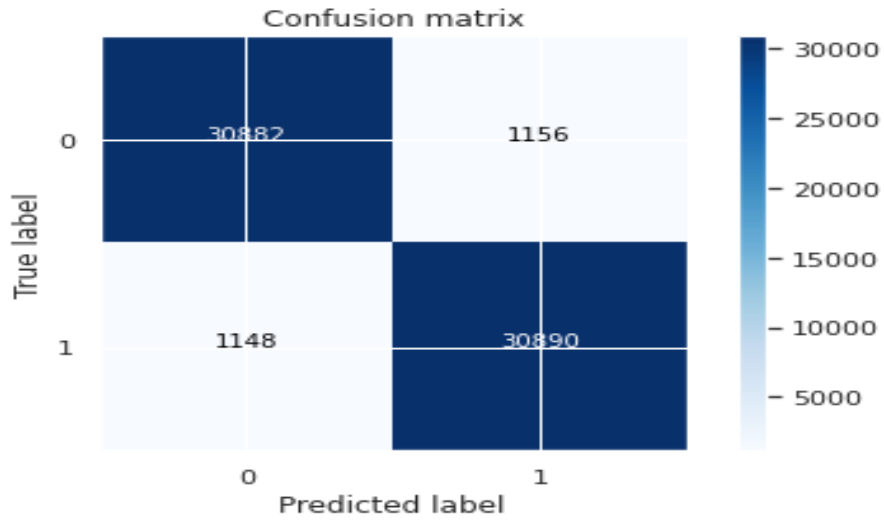


Figure 30: Confusion Matrix of Random Forest in Combined Training Set A and B

4.5 Classification Report:

Training Set A:

Xgboost:

	Precision	Recall	F1 Score	Support
0	0.84	0.95	0.89	156764
1	0.94	0.82	0.88	156764
accuracy			0.88	313528
macro avg	0.89	0.88	0.88	313528
weighted avg	0.89	0.88	0.88	313528

Figure 31: Classification Report of Xgboost in Training Set A

LightGBM

	Precision	Recall	F1 Score	Support
0	0.81	0.91	0.86	156764
1	0.90	0.78	0.84	156764
accuracy			0.85	313528
macro avg	0.85	0.85	0.85	313528
weighted avg	0.85	0.85	0.85	313528

:

Figure 32: Classification Report of LightGBM in Training Set A

RandomForest:

	Precision	Recall	F1 Score	Support
0	0.93	0.95	0.94	156764
1	0.95	0.92	0.94	156764
accuracy			0.94	313528
macro avg	0.94	0.94	0.94	313528
weighted avg	0.94	0.94	0.94	313528

Figure 33: Classification Report of RandomForest in Training Set A**Training Set B:****Xgboost:**

	Precision	Recall	F1 Score	Support
0	0.88	0.93	0.91	151606
1	0.93	0.87	0.90	151605
accuracy			0.90	303211
macro avg	0.90	0.90	0.90	303211
weighted avg	0.90	0.90	0.90	303211

Figure 34: Classification Report of Xgboost in Training Set B

LightGBM:

	Precision	Recall	F1 Score	Support
0	0.84	0.89	0.86	151606
1	0.88	0.83	0.85	151605
accuracy			0.86	303211
macro avg	0.86	0.86	0.86	303211
weighted avg	0.86	0.86	0.86	303211

Figure 35: Classification Report of LightGBM in Training Set B**RandomForest:**

	Precision	Recall	F1 Score	Support
0	0.96	0.96	0.96	151606
1	0.96	0.96	0.96	151605
accuracy			0.96	303211
macro avg	0.96	0.96	0.96	303211
weighted avg	0.96	0.96	0.96	303211

Figure 36: Classification Report of RandomForest in Training Set B

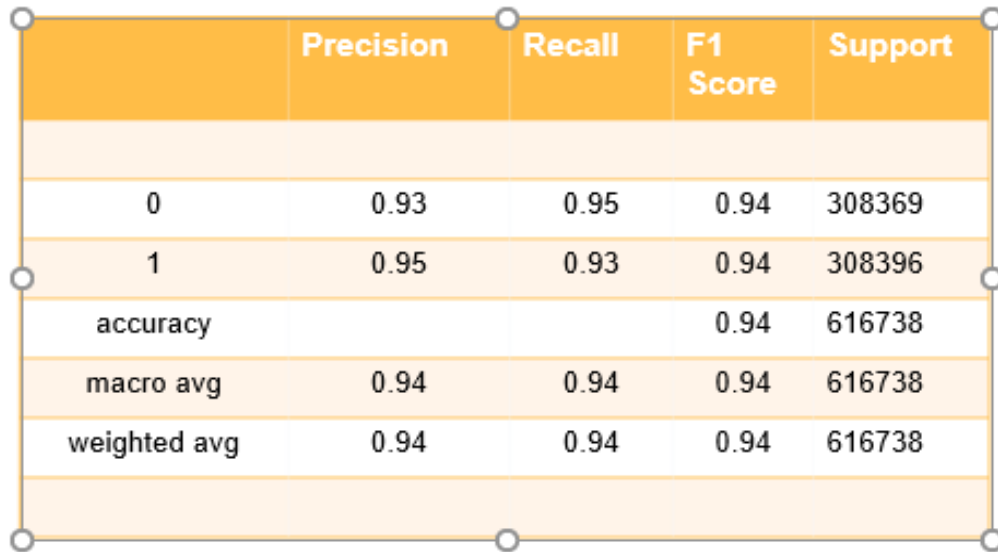
Training Set AB:**Xgboost:**

	Precision	Recall	F1 Score	Support
0	0.83	0.94	0.88	308369
1	0.93	0.81	0.87	308396
accuracy			0.88	616738
macro avg	0.88	0.88	0.88	616738
weighted avg	0.88	0.88	0.88	616738

Figure 37: Classification Report of Xgboost in Combined Training Set A and B**LightGBM:**

	Precision	Recall	F1 Score	Support
0	0.81	0.90	0.85	308369
1	0.89	0.79	0.84	308396
accuracy			0.85	616738
macro avg	0.85	0.85	0.84	616738
weighted avg	0.85	0.85	0.84	616738

Figure 38: Classification Report of LightGBM in Combined Training Set A and B

Random Forest:

	Precision	Recall	F1 Score	Support
0	0.93	0.95	0.94	308369
1	0.95	0.93	0.94	308396
accuracy			0.94	616738
macro avg	0.94	0.94	0.94	616738
weighted avg	0.94	0.94	0.94	616738

Figure 39: Classification Report of RandomForest in Combined Training Set A and B

4.6 Roc Curve:

Training Set A:

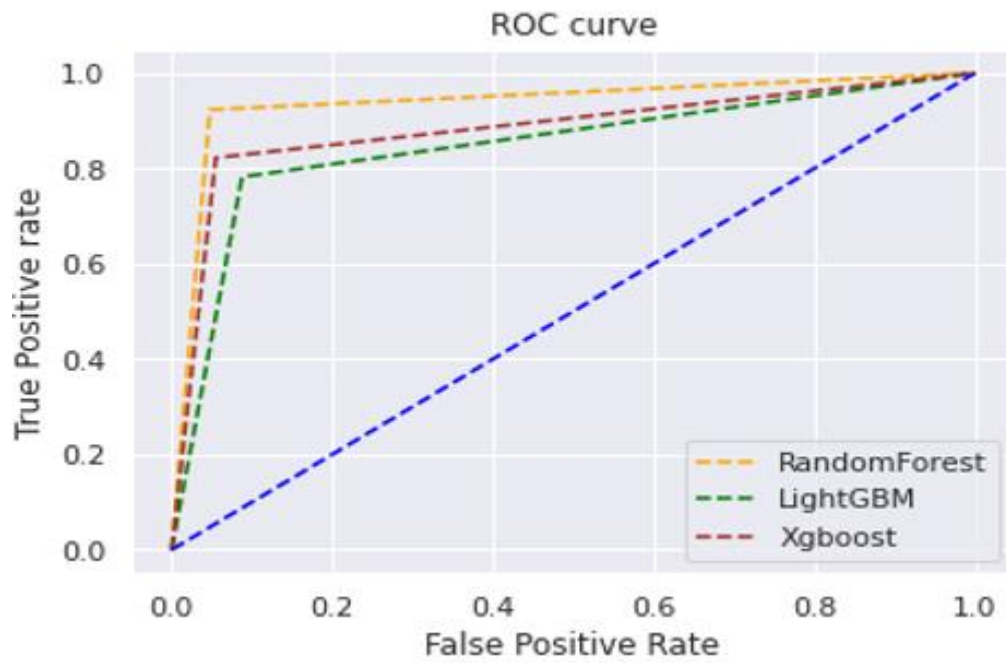


Figure 40: ROC curve of Training Set A

Training Set B:

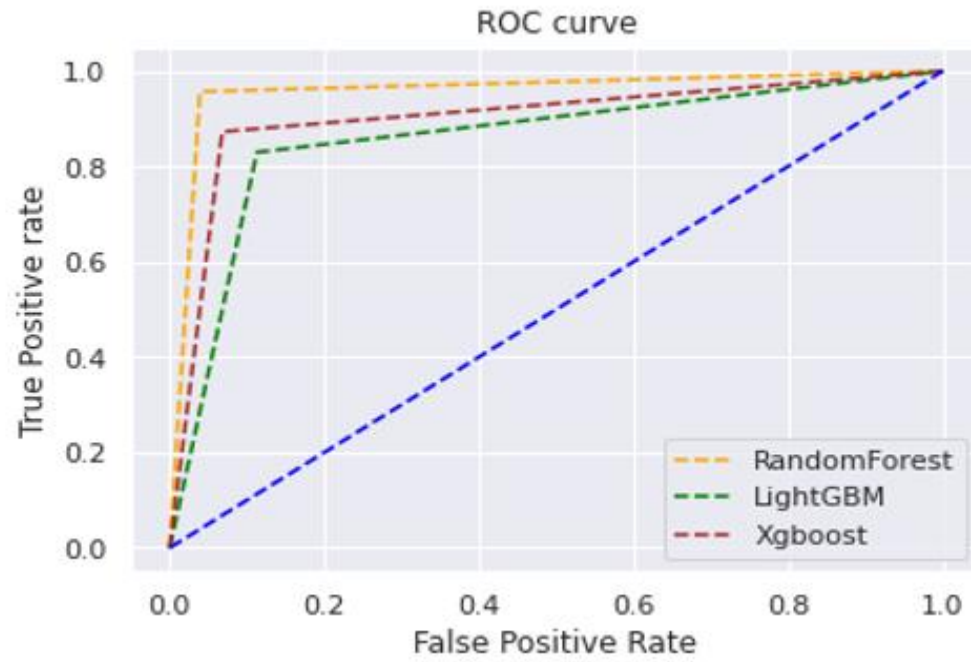


Figure 41: ROC curve of Training Set B

Training Set AB:

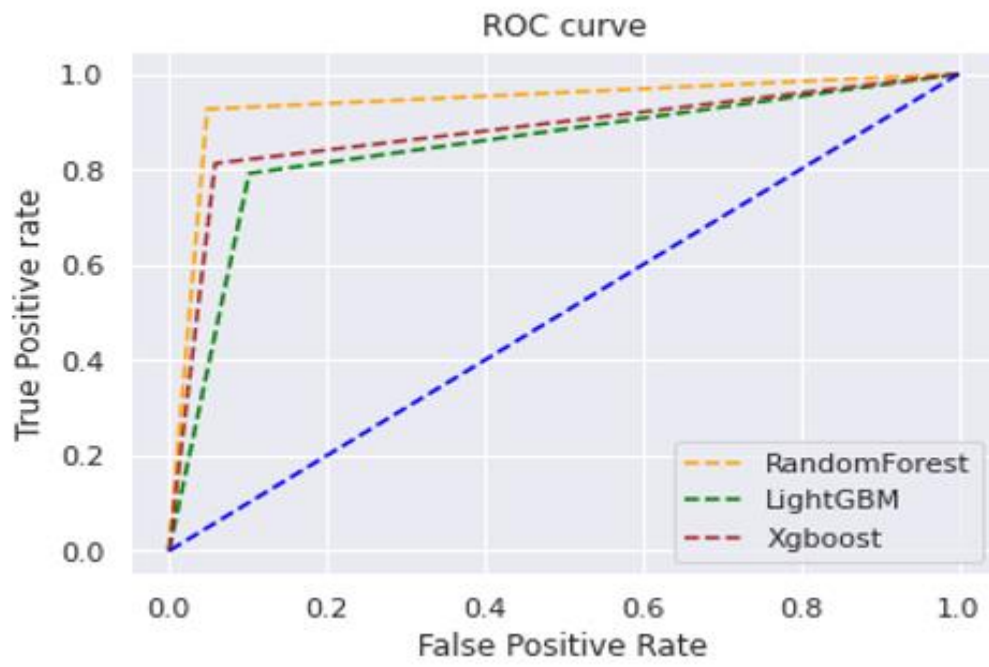


Figure 42: ROC curve of Combined Training Set A and B

Chapter 5

Discussion:

Early sepsis prediction is significant problem but still challenging. This study proposed that machine learning models shows high performance on prediction (ROC curve max 0.96) at the spot after patient's data entry (Figure 40,41,42). Machine learning algorithms used hourly based data after patients admitted in ICU to predict the prognosis of sepsis patients, the severity in condition of sepsis (i.e., septic shock), and maximum length of stay of septic patients in ICU. Xgboost, Random Forest and Lightgbm, classifiers had stronger predictive power, with areas under the ROC score of 0.90, 0.92,0.94 respectively. In early stage of sepsis, usage of Random Forest classifier allows to anticipate better ICU patient's outcome, shows appropriate medical measures and improve the treatment which improves prognosis.

As many biological events has happened in the pathophysiological of sepsis which leads to the disease processes and health complications. It's quite difficult to deal with disease complexity in ICU and imbalance data, therefore, the advanced methods of machine learning presented the new scoring systems for accurate prediction (Figure 1).

The another interesting outcome is every model trained on combined dataset Training set A and Training set B as well as on separate datasets and showing better results on training as well as on test dataset. Moreover, this study also shows the importance of each feature that is having great impact on sepsis. The statistical analysis has been used for the purpose of validation of each attribute based on Z-test. The total number of septic and non-septic patients in dataset are examined (Figure 19,20,21) and separate them in different classes and count the number of male and female having sepsis (Figure 13,14,15) and analyze the age which is more targeted due to sepsis (Figure 16,17,18). The prevalence of sepsis is disproportionately higher in the elder patients and the age of a person is an independent predictor of death. The elder patients are mostly non survivors of sepsis. This analysis is good for better understanding about the data and helpful to know that sepsis mostly effects

the female as compared to male. The difference in male and female shows different hormone response to an infection. The septic male and female have high estrogen level and shows the severity of illness in females than males. Females with septic shock have high anti-inflammatory mediators while males have high tendency to maintain the health status. So, by knowing the biological events it proved that females have severe effect towards illness than male.

After the statistical and correlation analysis six vital signs has confirmed (Figure 10,11,12) for the further process which are heart rate, temperature, oxygen saturation, respiratory rate, mean arterial pressure and systolic blood pressure. These variables having great impact in the prediction of sepsis and can be used for model building.

This study shows the contribution in the comparison of different machine learning models and find out the best models which can be deployed in hospitals. The model is trained on the features selected from dataset. For the prediction of sepsis, every model has presented best performance by giving ROC curve from (0.89 to 0.96). There is no limitation in distribution of features while using these models therefore, they can used to tackle the large data as well. The evaluation of predictive model occurs by confusion matrix which compute the sensitivity, error rate, precision and specificity while AUC is metric which differentiate the sepsis patients from other patients. In the comparison of these ensemble models, Random forest is more preferable than Xgboost because random forest is showing best precision and recall score as compared to Xgboost but Xgboost shows the integration of decision tress in sequential manner while random forest select each decision tree individually and make a random subset for construction (Figure 33,36,39). Every model could achieve highest ROC curve because of better selection of features, dealing with imbalance data or overfitting through smote analysis was the main key for the best prediction. Before SMOTE analysis the precision, recall and accuracy rate were very low. After implementation of SMOTE analysis, models showed best performance by using balanced data and predicted large number of true positives (sepsis patients are correctly identified as septic) and true negatives (non-sepsis patients are correctly identified as non-septic).

Conclusion:

Sepsis is life threatening disease which cause of high mortality rate and morbidity due to its ambiguous symptoms. Early detection is a key to overcome the death rate, therefore this study showed the development of fast and accurate machine learning algorithms Xgboost, Random Forest and LightGBM for the prediction of sepsis which give better results in form of ROC score from 0.90 to 0.96 .than the existing scoring systems i.e., SIRS, qSOFA, NEWS etc. In addition, the comparative analysis has done between five main models of machine learning by measuring their speed and their specificity and sensitivity range from 0.79-0.96. These models have potential to use for commercial use in ICUs for sepsis prediction.

References:

- Acuña, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. *Classification, Clustering, and Data Mining Applications*, 639–647.
- Adegbite, B., Edoa, J., EClinicalMedicine, W. N.-, & 2021, undefined. (n.d.). A comparison of different scores for diagnosis and mortality prediction of adults with sepsis in Low-and-Middle-Income Countries: a systematic review and. Elsevier. Retrieved January 5, 2022.
- Aljuaid, T., (ICDSE), S. S. D. S. and E., & 2016, undefined. (n.d.). Proper imputation techniques for missing values in data sets. *Ieeexplore.Ieee.Org*. Retrieved January 5, 2022.
- Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2018). Deep learning for missing value imputation in tables with non-numerical data. *International Conference on Information and Knowledge Management, Proceedings, 2017–2026*.
- Brink, A., Alsmá, J., Verdonshot, R. J. C. G., Rood, P. P. M., Zietse, R., Lingsma, H. F., & Schuit, S. C. E. (2019). Predicting mortality in patients with suspected sepsis at the Emergency Department; A retrospective cohort study comparing qSOFA, SIRS and National Early Warning Score. *PLoS ONE*, 14(1).
- Burdick, H., Pino, E., Gabel-Comeau, D., Gu, C., Roberts, J., Le, S., Slote, J., Saber, N., Pellegrini, E., Green-Saxena, A., Hoffman, J., & Das, R. (2020). Validation of a machine learning algorithm for early severe sepsis prediction: a retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals. *BMC Medical Informatics and Decision Making*, 20(1). <https://doi.org/10.1186/S12911-020-01284-X>
- Calvert, J., Price, D., Chettipally, U., ... C. B.-C. in biology, & 2016, undefined. (n.d.). A computational approach to early sepsis detection. Elsevier. Retrieved January 5, 2022,

- Chami, S., (CinC), K. T.-2019 C. in C., & 2019, undefined. (n.d.). Early Prediction of Sepsis From Clinical Data Using Single Light-GBM Model.
- CHAMI, S., Kaabouch, N., & Tavakolian, K. (n.d.). Comparative Study of Light-GBM and a Combination of Survival Analysis with Deep Learning for Early Detection of Sepsis.
- Chibani, S., & Coudert, F. X. (2020). Machine learning approaches for the prediction of materials properties. *APL Materials*, 8(8), 080701. <https://doi.org/10.1063/5.0018384>
- com, R. N.-D. uptodate., & 2008, undefined. (n.d.). Pathophysiology of sepsis. Do.Rsmu.Ru. Retrieved January 5, 2022
- Eachempati, S., Hydo, L., Surgery, P. B.-A. of, & 1999, undefined. (n.d.). Gender-based differences in outcome in patients with sepsis. *Jamanetwork.Com*. Retrieved January 5, 2022.
- Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14–15), 2627–2636.
- Goeij, M. de, Diepen, M. van, ... K. J.-N. D., & 2013, undefined. (n.d.). Multiple imputation: dealing with missing data. *Academic.Oup.Com*. Retrieved January 5, 2022.
- Goh, K., Wang, L., Yeow, A., Poh, H., ... K. L.-N., & 2021, undefined. (n.d.). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature.Com*. Retrieved January 5, 2022.
- Heidari, E., Sobati, M. A., & Movahedirad, S. (2016). Accurate prediction of nanofluid viscosity using a multilayer perceptron artificial neural network (MLP-ANN). *Chemometrics and Intelligent Laboratory Systems*.
- Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., & Wang, K. (2020). Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *Journal of Translational Medicine*, 18.

- Hsu, P., (CinC), C. H.-2019 C. in C., & 2019, undefined. (n.d.). A comparison of machine learning tools for early prediction of sepsis from icu data.
- Islam, M., Nasrin, T., Walther, B., ... C. W.-C. methods and, & 2019, undefined. (n.d.). Prediction of sepsis patients using machine learning approach: a meta-analysis. Elsevier. Retrieved January 5, 2022.
- Jang, J., Choi, J., Roh, H., ... S. S.-J. mHealth and, & 2020, undefined. (n.d.). Deep Learning Approach for Imputation of Missing Values in Actigraphy Data: Algorithm Development Study. Mhealth.Jmir.Org. Retrieved January 5, 2022.
- Kong, G., Lin, K., & Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Medical Informatics and Decision Making*, 20(1).
- Li, X., Xu, X., Xie, F., Xu, X., Sun, Y., Liu, X., ... X. J.-C. C., & 2020, undefined. (n.d.). A time-phased machine learning model for real-time prediction of sepsis in critical care. *Journals.Lww.Com*. Retrieved January 5, 2022.
- Liu, S., Ong, M., Mun, K., ... J. Y.-2019 C. in, & 2019, undefined. (n.d.). Early Prediction of Sepsis via SMOTE Upsampling and Mutual Information Based Downsampling.
- Mao, Q, Jay, M., Hoffman, J., Calvert, J., open, C. B.-B., & 2018, undefined. (n.d.). Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *Bmjopen.Bmj.Com*. Retrieved January 5, 2022
- Mao, Qingqing, Jay, M., Hoffman, J. L., Calvert, J., Barton, C., Shimabukuro, D., Shieh, L., Chettipally, U., Fletcher, G., Kerem, Y., Zhou, Y., & Das, R. (2018). Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU.
- Montomoli, J., Romeo, L., Moccia, S., Bernardini, M., Migliorelli, L., Berardini, D., Donati, A., Carsetti, A., Bocci, M. G., Wendel Garcia, P. D., Fumeaux, T., Guerci, P., Schüpbach, R. A., Ince, C., Frontoni, E., Hilty, M. P., Alfaro-Farias, M., Vizmanos-Lamotte, G., Tschöellitsch, T., ... Colak, E. (2021). Machine learning using the

- extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *Journal of Intensive Medicine*, 1(2), 110–116. <https://doi.org/10.1016/J.JOINTM.2021.09.002>
- Moor, M., Rieck, B., Horn, M., Jutzeler, C. R., & Borgwardt, K. (2021). Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. *Frontiers in Medicine*, 8, 348.
- Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU.
- Nesaragi, N., Sepsis, S. P.-I. D. and, & 2021, undefined. (n.d.). An Explainable Machine Learning Model for Early Prediction of Sepsis Using ICU Data. *Intechopen.Com*. Retrieved January 5, 2022
- Neviere, R., Parsons, P., Wolters, G. F.-M. en I., & 2017, undefined. (n.d.). Sepsis syndromes in adults: Epidemiology, definitions, clinical presentation, diagnosis, and prognosis. *Uptodate.Com*. Retrieved January 5, 2022
- O'Brien, J. M., Ali, N. A., Aberegg, S. K., & Abraham, E. (2007). Sepsis. *The American Journal of Medicine*, 120(12)
- Orhan, U., Hekim, M., & Ozer, M. (2011). EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, 38(10), 13475–13481.
- Pierrakos, C., & Vincent, J. L. (2010). Sepsis biomarkers: A review. *Critical Care*, 14(1), 1–18.
- Shenoy, K. V. V. (2020). Early Sepsis Prediction in Intensive Care Patients using Random Forest Classifier.
- Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J., & Das, R. (2017). 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*, 4(1), e000234.

- Singer, M., Deutschman, C., Jama, C. S., & 2016, undefined. (n.d.). The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jamanetwork.Com. Retrieved January 5, 2022,
- Stekhoven, D., Bioinformatics, P. B., & 2012, undefined. (n.d.). MissForest—non-parametric missing value imputation for mixed-type data. Academic.Oup.Com. Retrieved January 5, 2022.
- Su, L., Xu, Z., Chang, F., Ma, Y., Liu, S., Jiang, H., Wang, H., Li, D., Chen, H., Zhou, X., Hong, N., Zhu, W., & Long, Y. (2021). Early Prediction of Mortality, Severity, and Length of Stay in the Intensive Care Unit of Sepsis Patients Based on Sepsis 3.0 by Machine Learning Models. *Frontiers in Medicine*, 8.
- Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., & Hall, M. K. (2016). Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach. *Academic Emergency Medicine*, 23(3), 269–278.
- Usman, O., Usman, A., emergency, M. W.-T. A. journal of, & 2019, undefined. (n.d.). Comparison of SIRS, qSOFA, and NEWS for the early identification of sepsis in the Emergency Department. Elsevier. Retrieved January 7, 2022.
- van Doorn, W. P. T. M., Stassen, P. M., Borggreve, H. F., Schalkwijk, M. J., Stoffers, J., Bekers, O., & Meex, S. J. R. (2021). A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PLoS ONE*, 16(1 January).
- Wang, Dehua, Zhang, Y., & Zhao, Y. (2017). LightGBM: An effective miRNA classification method in breast cancer patients. *ACM International Conference Proceeding Series*, 7–11.
- Wang, Dong, Li, J., Sun, Y., Zhang, X., Liu, S., Han, B., & Wang, H. (2021). A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients in China.
- Wu, X., Akbarzadeh Khorshidi, H., Aickelin, U., Edib, Z., & Peate, M. (2019). Imputation techniques on missing values in breast cancer treatment and fertility data. *Health*

- Information Science and Systems, 7(1).
- Yang, M., Wang, X., Gao, H., Li, Y., Liu, X., Li, J., & Liu, C. (n.d.). Early prediction of sepsis using multi-feature fusion based XGBoost learning and Bayesian optimization.
- Yao, R. Q., Jin, X., Wang, G. W., Yu, Y., Wu, G. S., Zhu, Y. B., Li, L., Li, Y. X., Zhao, P. Y., Zhu, S. Y., Xia, Z. F., Ren, C., & Yao, Y. M. (2020). A Machine Learning-Based Prediction of Hospital Mortality in Patients With Postoperative Sepsis. *Frontiers in Medicine*, 7
- Zabihi, M., Kiranyaz, S., in, M. G.-2019 C., & 2019, undefined. (n.d.). Sepsis prediction in intensive care unit using ensemble of XGboost models.
- Zhao, X, Shen, W., and, G. W.-C. I., & 2021, undefined. (n.d.). Early Prediction of Sepsis Based on Machine Learning Algorithm. *Hindawi.Com*. Retrieved January 5, 2022,
- Zhao, Xuze, & Qu, B. (2021). A Comparative Study of Machine Learning Techniques for Predicting Sepsis for MIMIC-III Patients.
- Acuña, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. *Classification, Clustering, and Data Mining Applications*, 639–647.
- Adegbite, B., Edoa, J., *EClinicalMedicine*, W. N.-, & 2021, undefined. (n.d.). A comparison of different scores for diagnosis and mortality prediction of adults with sepsis in Low-and-Middle-Income Countries: a systematic review and. Elsevier. Retrieved January 5, 2022,
- Aljuaid, T., (ICDSE), S. S. D. S. and E., & 2016, undefined. (n.d.). Proper imputation techniques for missing values in data sets. *Ieeexplore.Ieee.Org*. Retrieved January 5, 2022,
- Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2018). Deep learning for missing value imputation in tables with non-numerical data. *International Conference on Information and Knowledge Management, Proceedings, 2017–2026*.
- Brink, A., Alsma, J., Verdonchot, R. J. C. G., Rood, P. P. M., Zietse, R., Lingsma, H. F.,

- & Schuit, S. C. E. (2019). Predicting mortality in patients with suspected sepsis at the Emergency Department; A retrospective cohort study comparing qSOFA, SIRS and National Early Warning Score. *PLoS ONE*, 14(1).
- Burdick, H., Pino, E., Gabel-Comeau, D., Gu, C., Roberts, J., Le, S., Slote, J., Saber, N., Pellegrini, E., Green-Saxena, A., Hoffman, J., & Das, R. (2020). Validation of a machine learning algorithm for early severe sepsis prediction: a retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals. *BMC Medical Informatics and Decision Making*, 20(1).
- Calvert, J., Price, D., Chettipally, U., ... C. B.-C. in biology, & 2016, undefined. (n.d.). A computational approach to early sepsis detection. Elsevier. Retrieved January 5, 2022,
- Chami, S., (CinC), K. T.-2019 C. in C., & 2019, undefined. (n.d.). Early Prediction of Sepsis From Clinical Data Using Single Light-GBM Model.
- CHAMI, S., Kaabouch, N., & Tavakolian, K. (n.d.). Comparative Study of Light-GBM and a Combination of Survival Analysis with Deep Learning for Early Detection of Sepsis.
- Chibani, S., & Coudert, F. X. (2020). Machine learning approaches for the prediction of materials properties. *APL Materials*, 8(8), 080701. , R. N.-D. uptodate., & 2008, undefined. (n.d.). Pathophysiology of sepsis. Do.Rsmu.Ru. Retrieved January 5, 2022,
- Eachempati, S., Hydo, L., Surgery, P. B.-A. of, & 1999, undefined. (n.d.). Gender-based differences in outcome in patients with sepsis. *Jamanetwork.Com*. Retrieved January 5, 2022.
- Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14–15), 2627–2636.
- Goeij, M. de, Diepen, M. van, ... K. J.-N. D., & 2013, undefined. (n.d.). Multiple imputation: dealing with missing data. *Academic.Oup.Com*. Retrieved January 5, 2022,

- Goh, K., Wang, L., Yeow, A., Poh, H., ... K. L.-N., & 2021, undefined. (n.d.). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. Nature.Com. Retrieved January 5, 2022,.
- Heidari, E., Sobati, M. A., & Movahedirad, S. (2016). Accurate prediction of nanofluid viscosity using a multilayer perceptron artificial neural network (MLP-ANN). *Chemometrics and Intelligent Laboratory Systems*.
- Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., & Wang, K. (2020). Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *Journal of Translational Medicine*, 18(1).
- Hsu, P., (CinC), C. H.-2019 C. in C., & 2019, undefined. (n.d.). A comparison of machine learning tools for early prediction of sepsis from icu data.
- Islam, M., Nasrin, T., Walther, B., ... C. W.-C. methods and, & 2019, undefined. (n.d.). Prediction of sepsis patients using machine learning approach: a meta-analysis. Elsevier. Retrieved January 5, 2022.
- Jang, J., Choi, J., Roh, H., ... S. S.-J. mHealth and, & 2020, undefined. (n.d.). Deep Learning Approach for Imputation of Missing Values in Actigraphy Data: Algorithm Development Study. *Mhealth.Jmir.Org*. Retrieved January 5, 2022.
- Kong, G., Lin, K., & Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Medical Informatics and Decision Making*, 20(1).
- Li, X., Xu, X., Xie, F., Xu, X., Sun, Y., Liu, X., ... X. J.-C. C., & 2020, undefined. (n.d.). A time-phased machine learning model for real-time prediction of sepsis in critical care. *Journals.Lww.Com*. Retrieved January 5, 2022, from https://journals.lww.com/ccmjournals/Fulltext/2020/10000/A_Time_Phased_Machine_Learning_Model_for_Real_Time.31.aspx
- Liu, S., Ong, M., Mun, K., ... J. Y.-2019 C. in, & 2019, undefined. (n.d.). Early Prediction of Sepsis via SMOTE Upsampling and Mutual Information Based Downsampling.
- Mao, Q, Jay, M., Hoffman, J., Calvert, J., open, C. B.-B., & 2018, undefined. (n.d.).

- Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *Bmjopen.Bmj.Com*. Retrieved January 5, 2022
- Mao, Qingqing, Jay, M., Hoffman, J. L., Calvert, J., Barton, C., Shimabukuro, D., Shieh, L., Chettipally, U., Fletcher, G., Kerem, Y., Zhou, Y., & Das, R. (2018). Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open*, 8(1), e017833.
- Montomoli, J., Romeo, L., Moccia, S., Bernardini, M., Migliorelli, L., Berardini, D., Donati, A., Carsetti, A., Bocci, M. G., Wendel Garcia, P. D., Fumeaux, T., Guerci, P., Schüpbach, R. A., Ince, C., Frontoni, E., Hilty, M. P., Alfaro-Farias, M., Vizmanos-Lamotte, G., Tschoellitsch, T., ... Colak, E. (2021). Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *Journal of Intensive Medicine*, 1(2), 110–116.
- Moor, M., Rieck, B., Horn, M., Jutzeler, C. R., & Borgwardt, K. (2021). Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. *Frontiers in Medicine*, 8, 348.
- Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical Care Medicine*, 46(4), 547.
- Nesaragi, N., Sepsis, S. P.-I. D. and, & 2021, undefined. (n.d.). An Explainable Machine Learning Model for Early Prediction of Sepsis Using ICU Data. *Intechopen.Com*. Retrieved January 5, 2022.
- Neviere, R., Parsons, P., Wolters, G. F.-M. en I., & 2017, undefined. (n.d.). Sepsis syndromes in adults: Epidemiology, definitions, clinical presentation, diagnosis, and prognosis. *Uptodate.Com*. Retrieved January 5, 2022.
- O'Brien, J. M., Ali, N. A., Aberegg, S. K., & Abraham, E. (2007). Sepsis. *The American Journal of Medicine*, 120(12), 1012–1022.

- Orhan, U., Hekim, M., & Ozer, M. (2011). EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, 38(10), 13475–13481.
- Pierrakos, C., & Vincent, J. L. (2010). Sepsis biomarkers: A review. *Critical Care*, 14(1),
- Shenoy, K. V. V. (2020). Early Sepsis Prediction in Intensive Care Patients using Random Forest Classifier.
- Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J., & Das, R. (2017). 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*, 4(1), e000234.
- Singer, M., Deutschman, C., Jama, C. S., & 2016, undefined. (n.d.). The third international consensus definitions for sepsis and septic shock (Sepsis-3). *Jamanetwork.Com*. Retrieved January 5, 2022.
- Stekhoven, D., Bioinformatics, P. B., & 2012, undefined. (n.d.). MissForest—non-parametric missing value imputation for mixed-type data. *Academic.Oup.Com*. Retrieved January 5, 2022.
- Su, L., Xu, Z., Chang, F., Ma, Y., Liu, S., Jiang, H., Wang, H., Li, D., Chen, H., Zhou, X., Hong, N., Zhu, W., & Long, Y. (2021). Early Prediction of Mortality, Severity, and Length of Stay in the Intensive Care Unit of Sepsis Patients Based on Sepsis 3.0 by Machine Learning Models. *Frontiers in Medicine*, 8.
- Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., & Hall, M. K. (2016). Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach. *Academic Emergency Medicine*, 23(3), 269–278.
- Usman, O., Usman, A., emergency, M. W.-T. A. journal of, & 2019, undefined. (n.d.). Comparison of SIRS, qSOFA, and NEWS for the early identification of sepsis in the Emergency Department. *Elsevier*. Retrieved January 7, 2022,
- van Doorn, W. P. T. M., Stassen, P. M., Borggreve, H. F., Schalkwijk, M. J., Stoffers, J.,

- Bekers, O., & Meex, S. J. R. (2021). A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PLoS ONE*, 16(1 January).
- Wang, Dehua, Zhang, Y., & Zhao, Y. (2017). LightGBM: An effective miRNA classification method in breast cancer patients. *ACM International Conference Proceeding Series*, 7–11.
- Wang, Dong, Li, J., Sun, Y., Zhang, X., Liu, S., Han, B., & Wang, H. (2021). A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients in China.
- Wu, X., Akbarzadeh Khorshidi, H., Aickelin, U., Edib, Z., & Peate, M. (2019). Imputation techniques on missing values in breast cancer treatment and fertility data. *Health Information Science and Systems*, 7(1).
- Yang, M., Wang, X., Gao, H., Li, Y., Liu, X., Li, J., & Liu, C. (n.d.). Early prediction of sepsis using multi-feature fusion based XGBoost learning and Bayesian optimization.
- Yao, R. Q., Jin, X., Wang, G. W., Yu, Y., Wu, G. S., Zhu, Y. B., Li, L., Li, Y. X., Zhao, P. Y., Zhu, S. Y., Xia, Z. F., Ren, C., & Yao, Y. M. (2020). A Machine Learning-Based Prediction of Hospital Mortality in Patients With Postoperative Sepsis. *Frontiers in Medicine*, 7.
- Zabihi, M., Kiranyaz, S., in, M. G.-2019 C., & 2019, undefined. (n.d.). Sepsis prediction in intensive care unit using ensemble of XGboost models.
- Zhao, X, Shen, W., and, G. W.-C. I., & 2021, undefined. (n.d.). Early Prediction of Sepsis Based on Machine Learning Algorithm. *Hindawi.Com*. Retrieved January 5, 2022,
- Zhao, Xuze, & Qu, B. (2021). A Comparative Study of Machine Learning Techniques for Predicting Sepsis for MIMIC-III Patients.
- Acuña, E., & Rodriguez, C. (2004). The Treatment of Missing Values and its Effect on Classifier Accuracy. *Classification, Clustering, and Data Mining Applications*, 639–647.
- Adegbite, B., Edoa, J., *EClinicalMedicine*, W. N.-, & 2021, undefined. (n.d.). A

- comparison of different scores for diagnosis and mortality prediction of adults with sepsis in Low-and-Middle-Income Countries: a systematic review and. Elsevier. Retrieved January 5, 2022,
- Aljuaid, T., (ICDSE), S. S. D. S. and E., & 2016, undefined. (n.d.). Proper imputation techniques for missing values in data sets. Ieeexplore.Ieee.Org. Retrieved January 5, 2022.
- Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., & Lange, D. (2018). Deep learning for missing value imputation in tables with non-numerical data. International Conference on Information and Knowledge Management, Proceedings, 2017–2026.
- Brink, A., Alsmas, J., Verdonchot, R. J. C. G., Rood, P. P. M., Zietse, R., Lingsma, H. F., & Schuit, S. C. E. (2019). Predicting mortality in patients with suspected sepsis at the Emergency Department; A retrospective cohort study comparing qSOFA, SIRS and National Early Warning Score. PLoS ONE, 14(1).
- Burdick, H., Pino, E., Gabel-Comeau, D., Gu, C., Roberts, J., Le, S., Slote, J., Saber, N., Pellegrini, E., Green-Saxena, A., Hoffman, J., & Das, R. (2020). Validation of a machine learning algorithm for early severe sepsis prediction: a retrospective study predicting severe sepsis up to 48 h in advance using a diverse dataset from 461 US hospitals. BMC Medical Informatics and Decision Making, 20(1).
- Calvert, J., Price, D., Chettipally, U., ... C. B.-C. in biology, & 2016, undefined. (n.d.). A computational approach to early sepsis detection. Elsevier. Retrieved January 5, 2022,
- Chami, S., (CinC), K. T.-2019 C. in C., & 2019, undefined. (n.d.). Early Prediction of Sepsis From Clinical Data Using Single Light-GBM Model..
- CHAMI, S., Kaabouch, N., & Tavakolian, K. (n.d.). Comparative Study of Light-GBM and a Combination of Survival Analysis with Deep Learning for Early Detection of Sepsis.
- Chibani, S., & Coudert, F. X. (2020). Machine learning approaches for the prediction of materials properties. APL Materials, 8(8), 080701.
- R. N.-D. uptodate., & 2008, undefined. (n.d.). Pathophysiology of sepsis. Do.Rsmu.Ru.

- Retrieved January 5, 2022.
- Eachempati, S., Hydo, L., Surgery, P. B.-A. of, & 1999, undefined. (n.d.). Gender-based differences in outcome in patients with sepsis. Jamanetwork.Com. Retrieved January 5, 2022.
- Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric Environment*,
- Goeij, M. de, Diepen, M. van, ... K. J.-N. D., & 2013, undefined. (n.d.). Multiple imputation: dealing with missing data. Academic.Oup.Com. Retrieved January 5, 2022,
- Goh, K., Wang, L., Yeow, A., Poh, H., ... K. L.-N., & 2021, undefined. (n.d.). Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature.Com*. Retrieved January 5, 2022,
- Heidari, E., Sobati, M. A., & Movahedirad, S. (2016). Accurate prediction of nanofluid viscosity using a multilayer perceptron artificial neural network (MLP-ANN). *Chemometrics and Intelligent Laboratory Systems*, 155, 73–85.
- Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., & Wang, K. (2020). Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *Journal of Translational Medicine*, 18(1).
- Hsu, P., (CinC), C. H.-2019 C. in C., & 2019, undefined. (n.d.). A comparison of machine learning tools for early prediction of sepsis from icu data.
- Islam, M., Nasrin, T., Walther, B., ... C. W.-C. methods and, & 2019, undefined. (n.d.). Prediction of sepsis patients using machine learning approach: a meta-analysis. Elsevier. Retrieved January 5, 2022,
- Jang, J., Choi, J., Roh, H., ... S. S.-J. mHealth and, & 2020, undefined. (n.d.). Deep Learning Approach for Imputation of Missing Values in Actigraphy Data: Algorithm Development Study. *Mhealth.Jmir.Org*. Retrieved January 5, 2022,

- Kong, G., Lin, K., & Hu, Y. (2020). Using machine learning methods to predict in-hospital mortality of sepsis patients in the ICU. *BMC Medical Informatics and Decision Making*, 20(1).
- Li, X., Xu, X., Xie, F., Xu, X., Sun, Y., Liu, X., ... X. J.-C. C., & 2020, undefined. (n.d.). A time-phased machine learning model for real-time prediction of sepsis in critical care. *Journals.Lww.Com*. Retrieved January 5, 2022,
- Liu, S., Ong, M., Mun, K., ... J. Y.-2019 C. in, & 2019, undefined. (n.d.). Early Prediction of Sepsis via SMOTE Upsampling and Mutual Information Based Downsampling.
- Mao, Q, Jay, M., Hoffman, J., Calvert, J., open, C. B.-B., & 2018, undefined. (n.d.). Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *Bmjopen.Bmj.Com*. Retrieved January 5, 2022,
- Mao, Qingqing, Jay, M., Hoffman, J. L., Calvert, J., Barton, C., Shimabukuro, D., Shieh, L., Chettipally, U., Fletcher, G., Kerem, Y., Zhou, Y., & Das, R. (2018). Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open*, 8(1),
- Montomoli, J., Romeo, L., Moccia, S., Bernardini, M., Migliorelli, L., Berardini, D., Donati, A., Carsetti, A., Bocci, M. G., Wendel Garcia. (2021). Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients. *Journal of Intensive Medicine*, 1(2), 110–116.
- Moor, M., Rieck, B., Horn, M., Jutzeler, C. R., & Borgwardt, K. (2021). Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. *Frontiers in Medicine*, 8, 348.
- Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical Care Medicine*, 46(4), 547.
- Nesaragi, N., Sepsis, S. P.-I. D. and, & 2021, undefined. (n.d.). An Explainable Machine

- Learning Model for Early Prediction of Sepsis Using ICU Data. Intechopen.Com. Retrieved January 5, 2022,
- Neviere, R., Parsons, P., Wolters, G. F.-M. en I., & 2017, undefined. (n.d.). Sepsis syndromes in adults: Epidemiology, definitions, clinical presentation, diagnosis, and prognosis. Uptodate.Com. Retrieved January 5, 2022
- O'Brien, J. M., Ali, N. A., Aberegg, S. K., & Abraham, E. (2007). Sepsis. *The American Journal of Medicine*, 120(12), 1012–1022.
- Orhan, U., Hekim, M., & Ozer, M. (2011). EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. *Expert Systems with Applications*, 38(10), 13475–13481.
- Pierrakos, C., & Vincent, J. L. (2010). Sepsis biomarkers: A review. *Critical Care*, 14(1), 1–18.
- Shenoy, K. V. V. (2020). Early Sepsis Prediction in Intensive Care Patients using Random Forest Classifier.
- Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J., & Das, R. (2017). 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*, 4(1), e000234.
- Singer, M., Deutschman, C., Jama, C. S.-, & 2016, undefined. (n.d.). The third international consensus definitions for sepsis and septic shock (Sepsis-3). Jamanetwork.Com. Retrieved January 5, 2022,
- Stekhoven, D., Bioinformatics, P. B.-, & 2012, undefined. (n.d.). MissForest—non-parametric missing value imputation for mixed-type data. Academic.Oup.Com. Retrieved January 5, 2022
- Su, L., Xu, Z., Chang, F., Ma, Y., Liu, S., Jiang, H., Wang, H., Li, D., Chen, H., Zhou, X., Hong, N., Zhu, W., & Long, Y. (2021). Early Prediction of Mortality, Severity, and Length of Stay in the Intensive Care Unit of Sepsis Patients Based on Sepsis 3.0 by Machine Learning Models. *Frontiers in Medicine*, 8.

- Taylor, R. A., Pare, J. R., Venkatesh, A. K., Mowafi, H., Melnick, E. R., Fleischman, W., & Hall, M. K. (2016). Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data–Driven, Machine Learning Approach. *Academic Emergency Medicine*, 23(3), 269–278.
- Usman, O., Usman, A., emergency, M. W.-T. A. journal of, & 2019, undefined. (n.d.). Comparison of SIRS, qSOFA, and NEWS for the early identification of sepsis in the Emergency Department. Elsevier. Retrieved January 7, 2022,
- van Doorn, W. P. T. M., Stassen, P. M., Borggreve, H. F., Schalkwijk, M. J., Stoffers, J., Bekers, O., & Meex, S. J. R. (2021). A comparison of machine learning models versus clinical evaluation for mortality prediction in patients with sepsis. *PLoS ONE*, 16(1 January).
- Wang, Dehua, Zhang, Y., & Zhao, Y. (2017). LightGBM: An effective miRNA classification method in breast cancer patients. *ACM International Conference Proceeding Series*, 7–11.
- Wang, Dong, Li, J., Sun, Y., Zhang, X., Liu, S., Han, B., & Wang, H. (2021). A Machine Learning Model for Accurate Prediction of Sepsis in ICU Patients in China.
- Wu, X., Akbarzadeh Khorshidi, H., Aickelin, U., Edib, Z., & Peate, M. (2019). Imputation techniques on missing values in breast cancer treatment and fertility data. *Health Information Science and Systems*, 7(1).
- Yang, M., Wang, X., Gao, H., Li, Y., Liu, X., Li, J., & Liu, C. (n.d.). Early prediction of sepsis using multi-feature fusion based XGBoost learning and Bayesian optimization. *Cinc.Org*
- Yao, R. Q., Jin, X., Wang, G. W., Yu, Y., Wu, G. S., Zhu, Y. B., Li, L., Li, Y. X., Zhao, P. Y., Zhu, S. Y., Xia, Z. F., Ren, C., & Yao, Y. M. (2020). A Machine Learning-Based Prediction of Hospital Mortality in Patients With Postoperative Sepsis. *Frontiers in Medicine*, 7.
- Zabihi, M., Kiranyaz, S., in, M. G.-2019 C., & 2019, undefined. (n.d.). Sepsis prediction in intensive care unit using ensemble of XGboost models.

Zhao, X, Shen, W., and, G. W.-C. I., & 2021, undefined. (n.d.). Early Prediction of Sepsis Based on Machine Learning Algorithm. Hindawi.Com. Retrieved January 5, 2022,

Zhao, Xuze, & Qu, B. (2021). A Comparative Study of Machine Learning Techniques for Predicting Sepsis for MIMIC-III Patients.