CALIFORNIA STATE UNIVERSITY, NORTHRIDGE

Minimizing Deep Incisional and Organ/Space Surgical site infections (SSIs) in California acute care hospitals using Predictive Risk Analysis

A thesis submitted in partial fulfillment of the requirements

For the degree of Master of Science in

Computer Science

By

Mayur Sanjiokumar Rahangdale

December 2022

The thesis of Mayur Rahangdale is approved.

Dr. Abhishek Verma	Date
Dr. Katya Mkrtchyan	Date
Dr. Taehyung Wang, Chair	Date

California State University, Northridge

Acknowledgements

I would like to sincerely thank my committee chair, Dr. Taehyung Wang, Professor, California State University, Northridge's College of Engineering and Computer Science. Professor, your unwavering support and confidence in me over the course of this thesis have been crucial for its completion.

I would like to express my gratitude towards my committee members, Dr. Abhishek Verma and Dr. Katya Mkrtchyan for meticulously reviewing my thesis and certifying my work.

This thesis would not have been possible without the knowledge imparted to me by all my professors throughout my Master's degree. I would thus like to thank them.

Lastly, I would express my sincere gratitude to my parents, who have always been the wind beneath my wings. Thank you for always believing in me even when I did not believe in myself. You have inspired me to excel in my academic career.

Table of Contents

Copyright....ii

Signature Page....iii

Acknowledgement....iv

List of Figures....vii

List of Tables....ix

Abstract....x

Chapter 1: Introduction....1

1.1 Objective....4

1.2 Problem Statement....5

Chapter 2: Random Forest and Previous work....7

2.1 Technical Approach....11

2.1.1 Exploratory Data Analysis....11

2.1.2 Predictive Data Analysis....12

Chapter 3: Dataset....13

Chapter 4: Experiments....15

4.1 EDA Methods....15

4.2 Data Manipulation....15

4.3 Model Definition and Training....18

Chapter 5: Results....22

5.1 Exploratory Data Analysis Results.....22

5.2 Predictive Data Analysis Results.....24

Chapter 6: Conclusion.....31

6.1 Limitations and Future work.....33

References.....35

List of Figures

Figure 1: Illustration of SSI types……3

Figure 2: Impact of SSIs on Medical costs……..4

Figure 3: Random Forest for binary classification ……...……..7

Figure 4: Process of Sepsis development in human body …….…….9

Figure 5: Description of geospatial dataset ……...11

Figure 6: Description of clinical dataset ……...……12

Figure 7: Features of Physionet Dataset ……..…….13

Figure 8: Graph representing Missingness in data……16

Figure 9: Imbalance in classes(left) and balanced classes(right)……..17

Figure 10: SMOTE class balancing…….,,,……..18

Figure 11: Feature extraction showing selected features……..19

Figure 12: Choropleth map of Infection ratio in US counties……...22

Figure 13: Choropleth map of Infections reported in US counties……...…23

Figure 14(a): Evaluation metrics for Model 1……..…….25

Figure 14(b): Graphs of ROC curve, PR curve and Confusion Matrices for Model 1…….…….26

Figure 15(a): Evaluation metrics for Model 2……..…….27

Figure 15(b): Graphs of ROC curve, PR curve and Confusion Matrices for Model 2…….…….28

Figure 16(a): Evaluation metrics for Model 3……..…….29

Figure 16(b): Graphs of ROC curve, PR curve and Confusion Matrices for Model 3….…….30

List of Tables

Table 1: SSI Categorization in human body……2

Table 2: Description of Machine Learning models……...20

Abstract

Minimizing Deep Incisional and Organ/Space Surgical site infections (SSIs) in California acute care hospitals using Predictive Risk Analysis

By

Mayur Sanjiokumar Rahangdale

Master of Science in Computer Science

In the US, adults typically struggle to cover their medical expenses. Adults with lower incomes and limited insurance options are more likely to acknowledge this, but even people with greater salaries and health insurance are not immune to the high cost of healthcare. It is very or somewhat difficult for nearly half the American adults (47%) to pay for their medical bills.

As the healthcare sector transitions from volume-based to value-based treatment, there is a significant change taking place. Cost containment is a crucial component of any healthcare organization's strategy as the pandemic continues to put stress on our healthcare system and operating margins have shrunk. Artificial intelligence (AI) and automation technologies are how healthcare providers deliver better, more affordable care in a sector with razor-thin margins.

The cost of providing care is being reduced, and simple administrative processes like patient registration, patient data entry, claims to process, and much more are being streamlined. However, merely controlling healthcare costs is no longer sufficient; hospitals must make

significant investments in workflow optimization to save operating expenses while enhancing patient care. In order to reduce costs, it is necessary to pinpoint areas where machine learning as well as artificial intelligence can be incorporated in the healthcare industry.

This study is based on one of the similar domains where ML is applied to automate the sepsis prediction for the patients who underwent surgery in California hospitals. Utilization management (UM) is essential for facilitating patients' access to quality healthcare. AI technologies have a significant opportunity to save healthcare costs and enhance the patient-provider interaction through the prior authorization procedure. Effective utilization management is crucial because it gives doctors the ability to make decisions at the point of treatment, allowing patients to receive the high-quality care they need and deserve more quickly.

This study focuses on providing machine learning solutions for hospitals, utilizing authorized sharing of patient medical data as a tool to reduce Surgical Site Infection (SSI) rate in healthcare facilities in the state of California. The goal of this study is to incorporate Utilization management and Machine learning tools to deliver cost-optimization solutions for SSI rate reduction which can be practically adaptable by healthcare facilities in California.

Chapter 1: Introduction

A natural defense mechanism against infections is our skin. Any surgery that results in a break in the skin can cause an infection, despite the numerous safety measures and protocols in place to prevent infection.[1] Surgical site infections (SSIs) is the medical term for these infections since they develop on the area of the body where the surgery was performed.

Beyond the intended outcome, surgery is a medical treatment that has a variety of physical effects on the patient. Any kind of operation puts the patient at risk for infection and other problems, some of which could lead to sepsis.[2]

SSIs often happen within 30 days of surgery, according to the CDC. Organ or space SSIs, Deep incisional SSIs, and Superficial incisional SSIs are the three forms of SSIs that are classified. The early signs of SSI, including fever, delayed healing, redness, discomfort, tenderness, warmth, and swelling, are caused by all forms of SSIs.[3] The distinction between different types of SSI based on symptoms and occurrence site is explained in Table 1.

SSI Type	Occurrence in body	Type-specific symptoms
Superficial incisional SSI	Area of skin over surgical site	May produce pus from the wound site
Deep incisional SSI	Beneath the incision area in muscle and the tissues surrounding the muscles	May produce pus inside the wound
Organ or space SSI	A body organ or a space between organs which is involved in surgery	May show discharge of pus from the skin into organ/body space may form abscess

Table 1. SSI Categorization in human body

Pus generation and abscess formation are frequent symptoms that appear in the latter stages of infection in all three SSI types.[4] The infection may have spread to damaged tissues or organs by the moment a pus development is discovered. Therefore, it is crucial to identify SSI in the initial stage itself to avoid further damage to the subject's body.

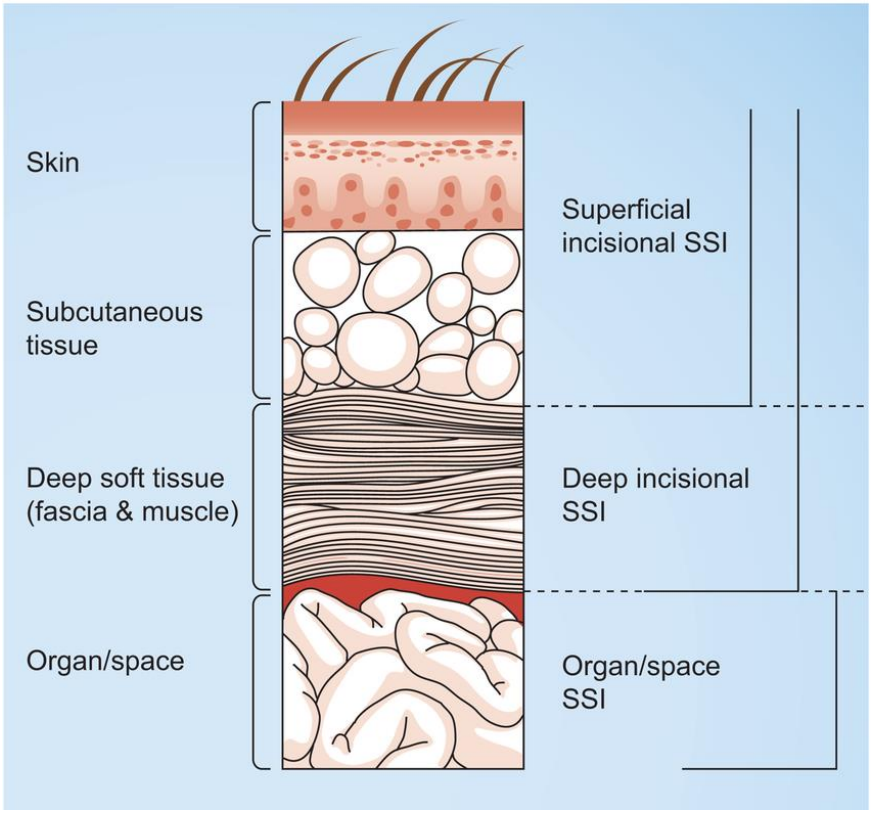

Fig. 1. Illustration of SSI types [4]

Sepsis is the body's response to an infection, which can cause organ failure, tissue damage, and even death., as illustrated in Fig. 1.[5] It is an aftereffect of many surgical site infections which can lead to severe sepsis or septic shock if not diagnosed early.[6] Early-stage Sepsis can be considered a vital key when identifying SSI in subjects that underwent surgery as sepsis occurrence can be used to trace back the chances of SSI from a former surgery that the subject underwent.[7]

This study proposes methodologies to minimize the SSI rate in California healthcare facilities by

1) Identifying SSI risk-prone medical facilities in California using historic SSI trends from Exploratory data-analysis methods.
2) Using Machine learning-based sepsis prediction in such facilities to identify sepsis in early-stage.

1.1 Objective

The main objective of this project is to identify risk-prone facilities for SSIs and suggest methodologies to be implemented in such facilities in order to minimize SSIs rates and their impact on the healthcare industry. The combined financial and human costs of tending surgical site infections increase exponentially,[8] involving direct and indirect costs such as extended hospitalization, additional dressings, more nursing care, and possible readmission to the hospital along with chances of further surgery.[9] This study is driven towards the minimization of the above impacts of SSI.

Fig. 2. Impact of SSIs on Medical costs [10]

Fig. 2 shows Early prediction of sepsis in patients that underwent surgery has a huge impact on mortality rate and medical costs. Over one-third of fatalities in American hospitals have sepsis, which affects nearly 1.7 million new cases and 270,000 deaths annually.[11] A majority of the $24 billion (13% of U.S. healthcare costs) that sepsis costs U.S. hospitals each year is for patients who were not initially diagnosed with the condition.[12] A prediction model for the early diagnosis of sepsis is put forth, and it makes use of clinical data collected from patients before, during, and after surgery in a hospital. The goal of this model is to analyze the patient profile and forecast the risk percentage of sepsis in a patient undergoing specific surgery after hospital discharge, following with suggesting preventive measures and follow-up health check-ups.

The two suggested approaches above covers the given problem as a whole. Given that about 50% of SSIs manifest after discharge, the number of SSIs is probably underestimated. Predicting chances of sepsis using patient feedback and clinical data, if or not a patient is at risk of post-surgery SSI, and whether or not requires readmission to the hospital.

1.2 Problem Statement

The goal of this research is to reduce the costs associated with sepsis, a serious illness that can result in organ failure, tissue damage, or even death. The goal of this study is to develop a working machine learning algorithm that can automatically detect a patient's risk of developing sepsis by using a positive or negative predictive model of sepsis for every timeframe centered on the clinical data provided. This study focuses on the early identification of sepsis utilizing physiological data.

The problem statement in this study can be divided into two significant parts. The first part is geospatial analysis of acute care hospitals in California to obtain regions/counties with spike in SSI rates for various deep Incisional and organ/space surgical site infections. The available dataset has a count of SSIs for every medical facility and the type of surgery. Based on this, geospatial maps and trends in SSI rates are derived for counties in the state of California.

The second part of problem statement is the prediction of chances of sepsis in a patient that underwent surgery. The data used in the prediction is the patient's demographic and clinical data like vital signs, laboratory values, and demographics during hospital admission. Based on this, a prediction model is derived to classify the patient into sepsis/non-sepsis classes. If any patient is classified into sepsis class, it implies that the patient has high chance of infection and may require either a follow-up or readmission to hospital.

A hypothesis is made that collectively the first half and second half of problem statement will act as an early sepsis prediction tool, hence cutting down the costs of SSI impact on human lives.

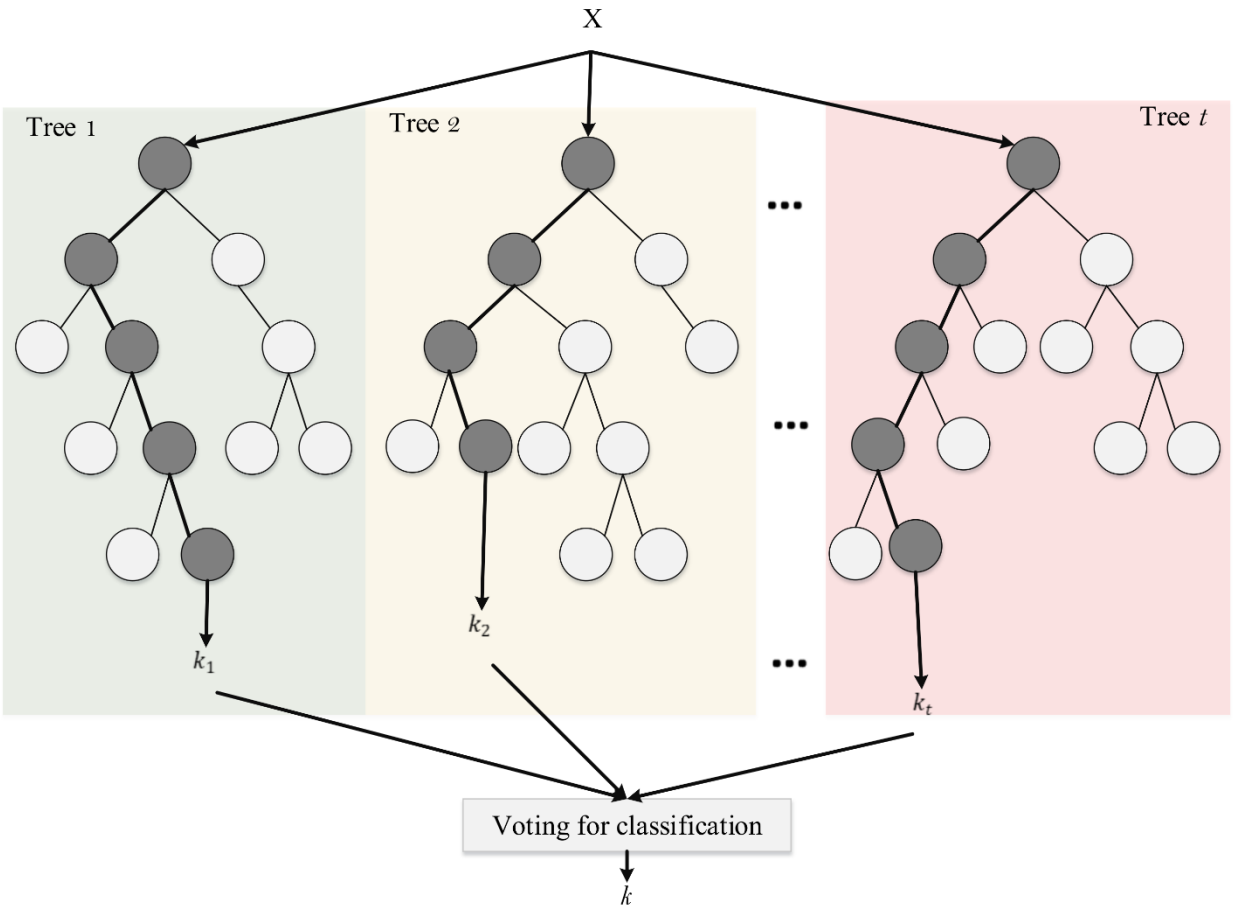

Fig. 3. Random forest for binary classification [15]

This study is aiming to achieve the task of binary classification by classifying the sepsis (true) labels and non-sepsis (false) labels. Using a randomly chosen portion of the training data, the random forest is a collection of decision trees. In Fig. 3, it shows that this algorithm operates as an ensemble where there is very low co-relation between the individual decision trees.[15] To select the final test object class, it combines the votes from various distinct decision trees. In this paper, the sepsis classification model is constructed using random forest.

The rapid Sequential Organ Failure Assessment (qSOFA) score was created utilizing multivariable logistic regression with a split sample for the assessment of "Clinical Criteria for Sepsis".[16] Its value as a prompt to investigate potential sepsis is supported by the fact that the predictive validity of the qSOFA for in-hospital mortality in this research was statically higher than score models in earlier studies. The introduced model [17] restricts the time window for primary analyses to only 24 hours after the onset of infection, whereas the score difference can occur even after 24 hours. A solution for this is required.

A majority of previous research papers have used singular binary classifiers like a neural network, logistic regression, and SVM and not implemented ensemble learning algorithms.[18,19,20] Considering the methodologies implemented in previous works, random forest is used as a choice of ensemble classifier for this study. This choice was made because, as long as they don't consistently all make the same blunder, the trees in the random forest shield one another from their particular flaws. As a result, it can be observed that a large number of decision trees that operate as a committee will perform better than any of their individual component models. There is a close relationship between surgical wounds and sepsis occurrences in the human body leading to surgical site infections which is explained in an article published in the Lancelet journal.[21]

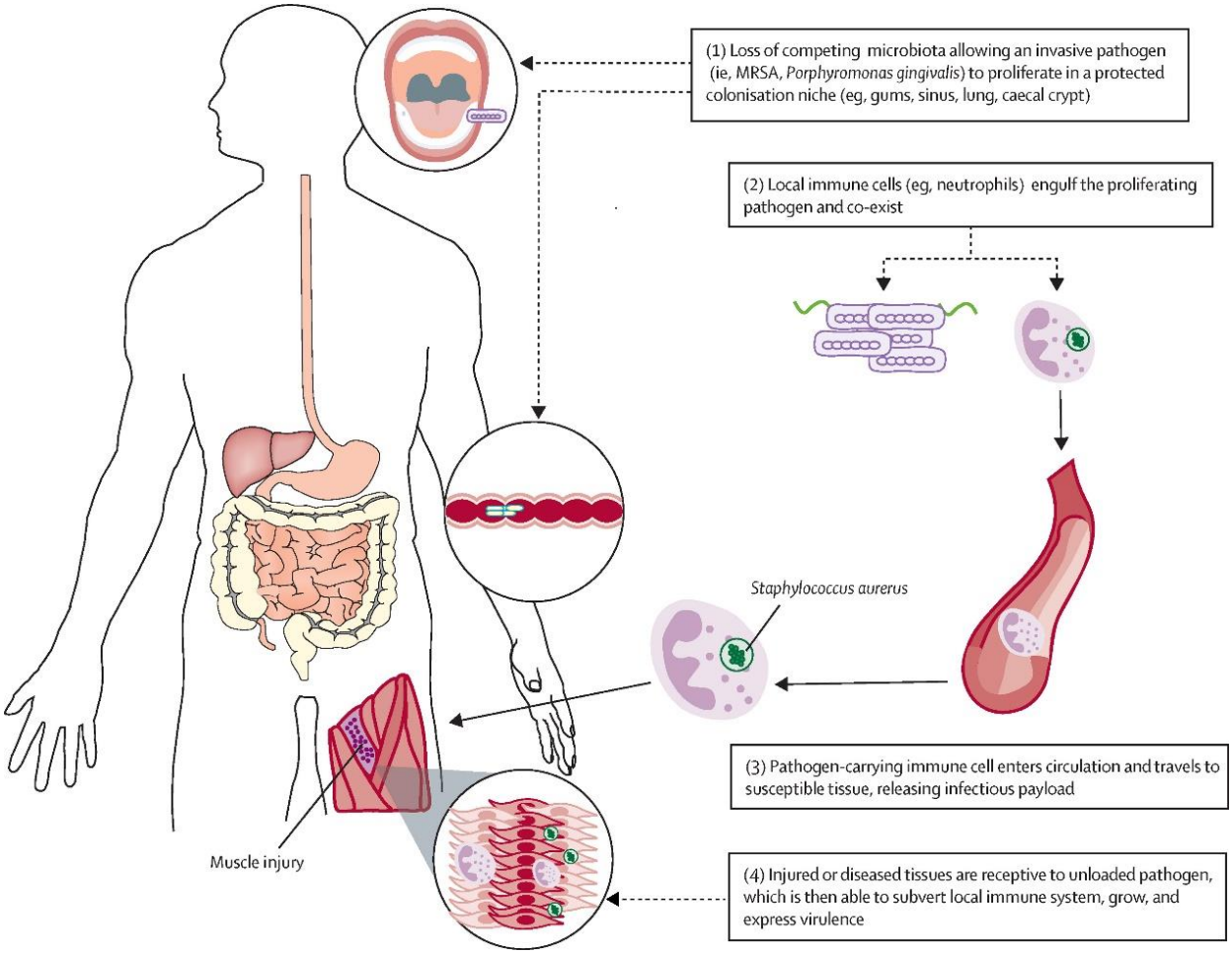

Fig. 4. Process of Sepsis development in human body [22]

Fig. 4. Shows the process of sepsis development in a human body when exposed to invasive pathogens. If any Invasive pathogens enter a human body and proliferate in a protected colonization niche (e.g., gums, sinus, lung, caecal crypt). As soon as these pathogens encounter immune cells, the local immune cells in these niches, the proliferating pathogens latch on to these local immune cells (e.g., neutrophils) and start co-existing in the blood stream. The pathogen carrying immune cells enter the blood circulation and they can circulate throughout the body.

There is a great chance that these cells travel to any susceptible tissue present in the body and release infectious payload. If the subject underwent surgery, there is a high chance that the susceptible tissues these cells attack are the injured/damaged tissues from the surgical site. These surgically induced disorders are extremely susceptible to other microorganisms, which can then proliferate and express virulence while undermining the local immune system. As a result, a sepsis infection grows over the surgical site, the body begins to display sepsis symptoms, and a surgical site infection result.[22]

Prior research[7] demonstrates that tracking vital signs of admitted patients is the first stage in creating a sepsis detection system, which has been demonstrated to be a successful solution to the issue of early sepsis identification. This study focuses on early sepsis detection based on vital signs as a tool to predict SSI waiting to occur at the surgical site of an individual who underwent surgery.

A hypothesis can be made that a greater time frame can be achieved to predict the ‘onset’ event using a more rigorous machine learning algorithm like the random forest algorithm. In addition to this, the use of more detailed gold standard datasets like the Physionet dataset which includes vital signs, laboratory values, and patient demographics can help achieve better sepsis label classification.

2.1 Technical Approach

The problem discussed above can be solved using exploratory data analysis methods and predictive modeling. Exploratory data analysis is the crucial process of doing preliminary analyses on data in order to identify anomalies, find patterns, test hypotheses, and double-check presumptions with the help of graphical representations and summary statistics. In this approach, EDA is applied on the hospitals' dataset to discover patterns in SSI occurrences in health-care facilities spread over the regions of California.

2.1.1 Exploratory Data Analysis

```
Index(['Year', 'State', 'County', 'HAI', 'Operative_Procedure', 'Facility_ID',
       'Facility_Name', 'Hospital_Category_RiskAdjustment', 'Facility_Type',
       'Procedure_Count', 'Infections_Reported', 'Infections_Predicted'],
      dtype='object')]
```

Fig. 5. Description of geospatial dataset

Fig. 5 shows all the columns of geospatial dataset including the locational column. Data manipulation using locational columns in geospatial dataset i.e., *State, County, Facility_Name* is done to obtain location-specific data using pandas and NumPy libraries. Based on this, county-specific SSI rates is calculated using NumPy operations. Choropleth maps are built using Geopandas and Geoplot libraries to analyze the count of reported infections and infection ratio in US counties. An overall geospatial data analysis of this dataset provides insights helpful in solving the first part of the problem.

2.1.2 Predictive Data Analysis

```
Index(['HR', 'O2Sat', 'Temp', 'SBP', 'MAP', 'DBP', 'Resp', 'EtCO2',
       'BaseExcess', 'HCO3', 'FiO2', 'pH', 'PaCO2', 'SaO2', 'AST', 'BUN',
       'Alkalinephos', 'Calcium', 'Chloride', 'Creatinine', 'Bilirubin_direct',
       'Glucose', 'Lactate', 'Magnesium', 'Phosphate', 'Potassium',
       'Bilirubin_total', 'TroponinI', 'Hct', 'Hgb', 'PTT', 'WBC',
       'Fibrinogen', 'Platelets', 'Age', 'Gender', 'Unit1', 'Unit2',
       'HospAdmTime', 'ICULOS', 'SepsisLabel'],
      dtype='object')
```

Fig. 6. Description of clinical dataset

The second part of the technical approach essentially duplicates the post-surgery scenario in the above medical facilities where the clinical data is collected from the duration of the patient's admission to discharge, as shown in Fig. 6. Utilization management is incorporated here, where hospitals are authorized to process medical data of admitted patients.

While working with both datasets to find a solution, it is assumed here that we have access to a patient's vital signs, laboratory values, and EHR data in the above-mentioned medical facilities. Patient demographics can be obtained through their anonymized EHR records.

A predictive machine learning model is built that feeds on vital features mentioned in Fig. 6 and predicts a *SepsisLabel* binary class value for the patient profile/row.

Chapter 3: Dataset

The dataset of surgical site infections (SSIs) used here is reported by California hospitals to the California Department of Public Health (CDPH), Healthcare-Associated Infections (HAI) Program, via the Centers for Disease Control and Prevention National Healthcare Safety Network (NHSN), in accordance with Health and Safety Code (HSC) section 1288.55.[13] The dataset for sepsis prediction is taken from 'Early Prediction of Sepsis from Clinical Data' -- the PhysioNet Computing in Cardiology Challenge 2019.[14] This dataset provides an opportunity to address the basic questions about the limitations of early sepsis detection which remains unanswered.

	HR	O2Sat	Temp	SBP	MAP	DBP	Resp	EtCO2	BaseExcess	HCO3	...	\
0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	
1	97.0	95.0	NaN	98.0	75.33	NaN	19.0	NaN	NaN	NaN	...	
2	89.0	99.0	NaN	122.0	86.00	NaN	22.0	NaN	NaN	NaN	...	
3	90.0	95.0	NaN	NaN	NaN	NaN	30.0	NaN	24.0	NaN	...	
4	103.0	88.5	NaN	122.0	91.33	NaN	24.5	NaN	NaN	NaN	...	

	WBC	Fibrinogen	Platelets	Age	Gender	Unit1	Unit2	HospAdmTime	\
0	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	
1	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	
2	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	
3	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	
4	NaN	NaN	NaN	83.14	0	NaN	NaN	-0.03	

	ICULOS	SepsisLabel
0	1	0
1	2	0
2	3	0
3	4	0
4	5	0

Fig. 7. Features of Physionet Dataset

Fig. 7 shows all the Feature categories of Physionet Dataset which are Vital signs, Laboratory values and patient demographics. Vital Signs are Heart Rate, Blood Pressure, Temperature, Respiratory rate, and End-tidal carbon dioxide. Laboratory Values are Platelet Count, Glucose, Calcium, etc. Patient demographics are Age, Gender, Time in ICU, and Hospital Admit time. Label values are 0 (non-sepsis) and 1 (Sepsis).

The data repository contains one file per patient (e.g., training/p00001.psv) for 40,000 patients. The table's columns each provide a series of measures over time. The table's rows each offer a set of measurements all at once. All rows are at a time difference of min 8 hours and a max of 2 weeks. Vital signs are recorded every hour and laboratory values are recorded every 24 hours.

Chapter 4: Experiments

4.1 EDA Methods

While importing the hospital dataset, the Geopandas library was used to merge the geographical coordinates of counties with the original dataset, out of which Alaska, Hawaii, and Puerto Rico were omitted from the merging. Using domain expertise, feature engineering was utilized to extract features (properties, characteristics, and attributes) from the raw data columns for further data preparation. The goal is to employ these extra attributes to enhance the quality of the results produced by the machine learning process as opposed to just giving it raw data. Counties were grouped by column sums of 'Procedure_Count', 'Infections_Reported', and 'Infections_Predicted'. A new feature ‘Infection_Ratio’ was engineered by operation ['Infections_Reported'] / ['Procedure_Count']. The final analysis was done by plotting choropleth maps using the Geoplot library, where, ‘Infections_Reported’ and ‘Infection_Ratio’ were geographically mapped on US counties.

4.2 Data Manipulation

While importing the sepsis dataset, PSV for each patient was concatenated into a single data frame and split into training, validation, and testing sets. The features with more than 50% missing values i.e. unit1, unit2, and EtCO2 were dropped from the dataset as they had no significant impact and could not act as important features.

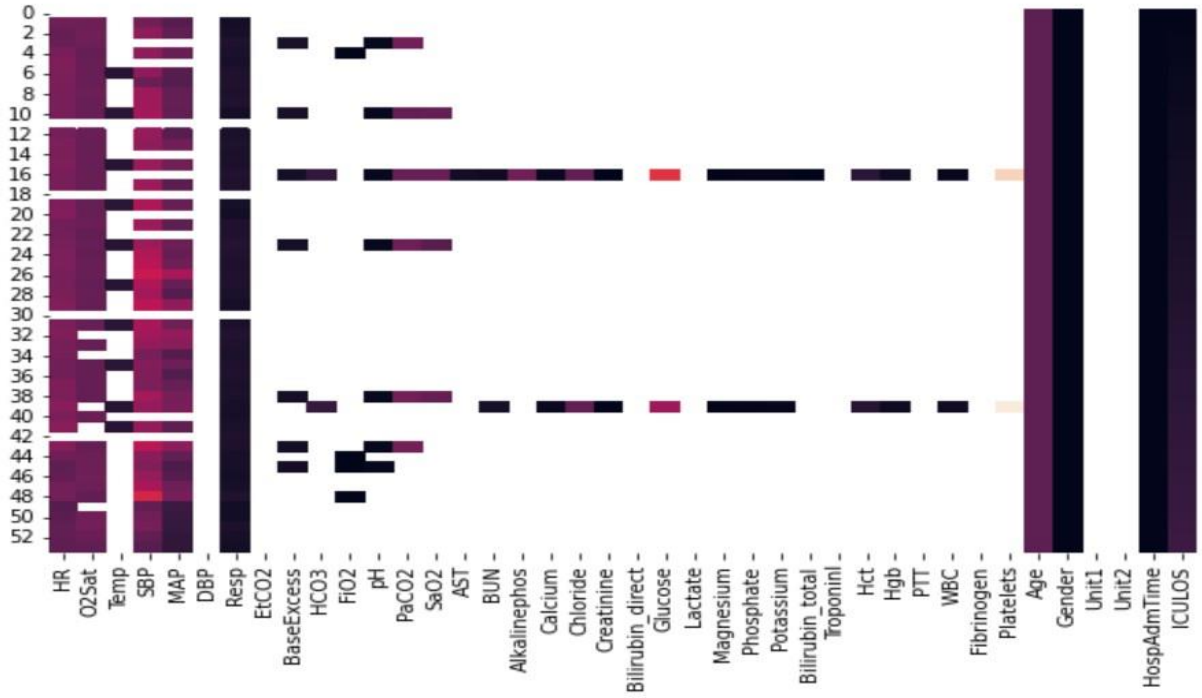

Fig. 8. Graph representing Missingness in data [23]

Due to the difference in recording frequency of vital signs and laboratory values, there were a significant amount of NaN values, as shows in Fig. 8. If missing data are not managed properly, they could substantially undermine the conclusions drawn from randomized clinical studies. The mechanisms leading the data to be absent, as well as the analytical techniques used to correct the missingness, both affect the possible bias brought on by missing data. Imputation, which substituted the NaN values in features by imputations, is used to tackle this missingness in the data.[23]

In the sepsis and non-sepsis classes, it was found during the examination of the sepsis label that there was a very considerable class imbalance.

Working with unbalanced datasets presents a difficulty because our machine learning model will overlook the minority class and perform poorly as a result. Class balancing is necessary since the minority class's performance is the most crucial.

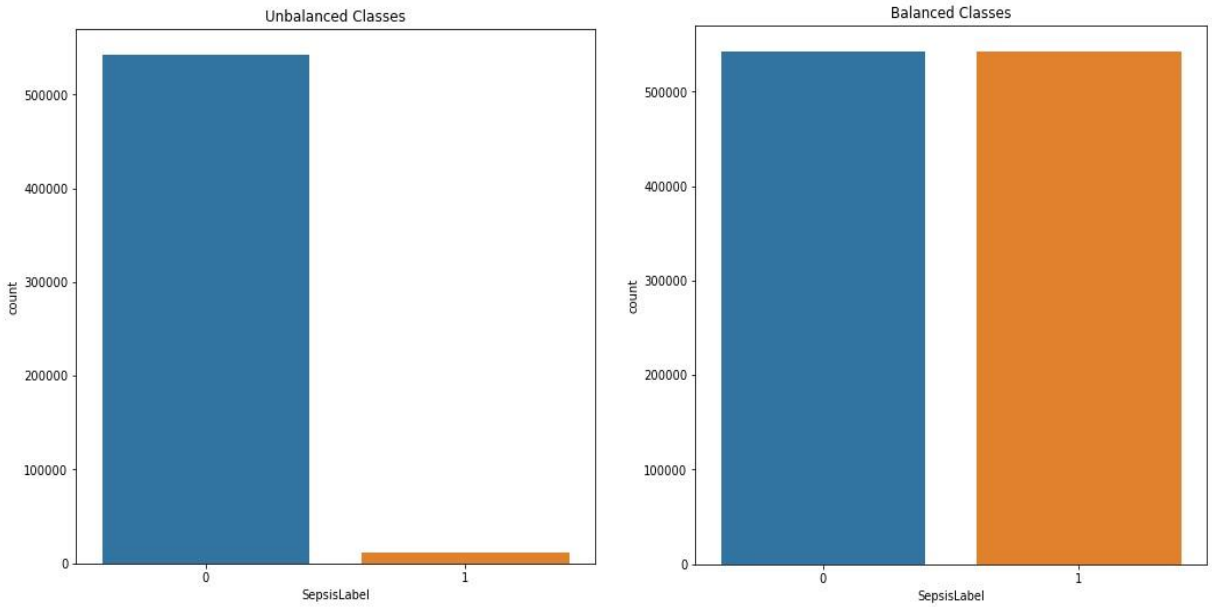

Fig. 9. Imbalance in classes(left) and balanced classes(right)

Synthetic Minority Oversampling Technique (SMOTE) is an oversampling method used to improve random oversampling. This method removes class imbalance in data for minority classes (true sepsis labels) using data augmentation, which can be illustrated in Fig. 10.[24]

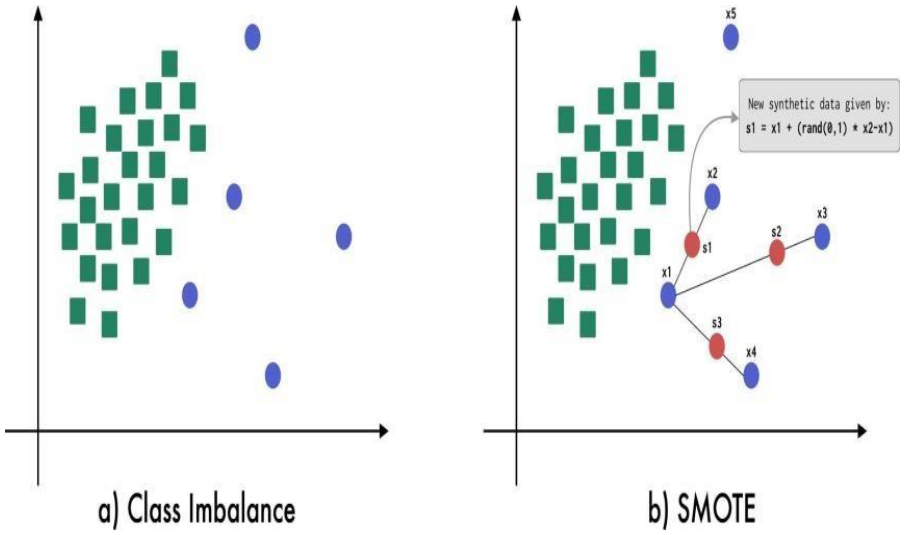

Fig. 10. SMOTE class balancing [24]

Newly synthesized true samples from SMOTE operation were added to the dataset to obtain a total of 1085606 samples. In the final data, the class imbalance is zero, as seen in Fig. 9.

4.3 Model Definition and Training

Random selection is used to make sample dataset of size 20,000 samples. On the sample dataset, the SelectFromModel function from the Sklearn feature selection module is used to select important features. After the feature selection step, 18 features were selected from the total columns present in the dataset, which are shown in Fig. 11. The features selected consisted of a majority of vital signs, a few of laboratory values and only a couple of numerical demographics such as age.

```
18 selected features

['HR',
 'Temp',
 'SBP',
 'MAP',
 'PaCO2',
 'BUN',
 'Creatinine',
 'Glucose',
 'Phosphate',
 'Potassium',
 'Hct',
 'Hgb',
 'PTT',
 'WBC',
 'Platelets',
 'Age',
 'HospAdmTime',
 'ICULOS']
```

Fig. 11. Feature extraction showing selected features

A total of three models were built with random forest classifier as the algorithm. The random forest classifiers were imported using Python's Sklearn library. The description of the three models can be found in Table. 2. The first model (Model 1) was trained without class balancing and feature selection. The second model (Model 2) was trained with feature selection but without class balancing. For the third model (Model 3), both class balancing and feature selection was used for training. All the models in the table are then evaluated by calculating values and plots for F1-score, precision, recall, PR, ROC curves, and accuracies.

Model name	Classifier	Description	Training size
Model 1	Random forest	All the features from the dataset used for training with class imbalance	20,000 samples
Model 2	Random forest	18 selected features used for training with class imbalance	20,000 samples
Model 3	Random forest	18 selected features used for training without class imbalance	20,000 samples

Table 2. Description of Machine Learning models

The models built above are imported as RandomForestClassifier from sklearn.ensemble library. The parameters for model definition (RF parameters) specified are n_estimators, random_state, oob_score, max_depth, class_weight, max_features, and verbose.

The count of individual trees in random forest model is set by n_estimators parameter. It's default value is 100 but it is set to 150 when defining our model. The maximum allowed depth of the tree is set by max_depth parameter. If the value of max_depth is set to None, then nodes are expanded till all leaves of tree are pure or until all leaves contain less than min_samples_split samples. The value of max_depth is set to 10. The number of features to consider when looking for the best split is set by max_features parameter.

The value of max_features is set to 37. The random_state parameter controls both the randomness of the bootstrapping of the samples used when building trees (if bootstrap=True) and the sampling of the features to consider when looking for the best split at each node (if max_features < n_features). The value of random_state parameter is set to 1211. The bootstrap parameter decides whether bootstrap samples are used when building trees. The value of bootstrap parameter is default which is True. The parameter oob_score decides whether to use out-of-bag samples to estimate the generalization score. It is only available if bootstrap=True, which is already set to default. The default value of oob_score is False but it is set to True. The value of verbose parameter is set to 2.

Chapter 5: Results

5.1 Exploratory Data Analysis Results

Fig. 12. Choropleth map of Infection ratio in US counties

The choropleth map in Fig. 12 depicts the numerical value of infection ratio in a color-coded manner, where darker regions represent high values of infection ratio and lighter regions represent lower values of infection ratio. Infection ratio represents SSI infection reported per surgical procedure from a medical facility in that county. In California state, San Diego, Kern, Imperial, and Ventura counties from the southern counties have high values of infection ratio. From the Bay Area, Alameda, Solano, Marin, and Contra Costa counties also show high values of infection ratio.

Sacramento and Fresno from the central counties and Shasta, Siskiyou, and Humboldt from superior counties also show considerably high values of infection ratio. On the other hand, the infection ratio is found low in the Los Angeles region and San Francisco.

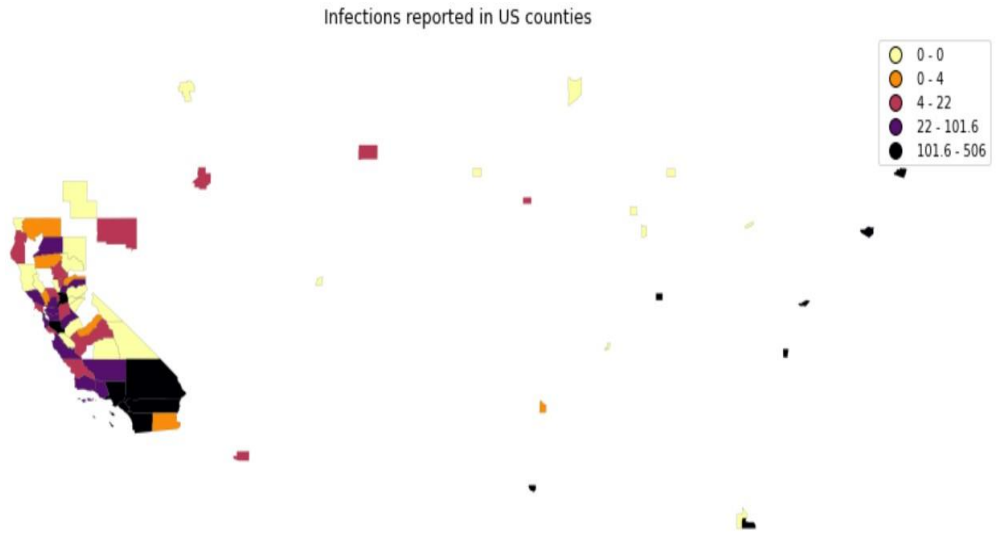

Fig. 13. Choropleth map of Infections reported in US counties

The choropleth map in Fig. 13 depicts the numerical count of infections reported in a color-coded manner, where darker regions represent high values of infection ratio and lighter regions represent lower values of infection ratio. The count of infections reported is the sum of the count of infections reported in medical facilities in that county. Southern counties like Ventura, Santa Barbara, Riverside, San Diego, San Bernardino, Orange, and Kern show high counts of infections reported. Almost all the southern counties show a high count of infections whereas the count is very low in central counties. From the Bay Area, counties such as Marin, Solano,

Alameda, Contra Costa, Santa Clara, and San Mateo show a high count of infections reported. Shasta county from superior counties and Los Angeles Region also shows a high count of infections reported.

5.2 Predictive Analysis Results

In this section, Fig. 14, Fig 15 and Fig.16 each represent a Precision-Recall Curve (PR Curve) and a confusion matrix mentioning precision, recall, f1-score and support. The best values of precision, recall, f1-score, Area under PR curve (AUPRC), Area under curve (AUC) and accuracy are listed below the matrix. The TPR-FPR graph has the ROC curve which represents Area under ROC (AUROC). Confusion matrix is illustrated in both normal and normalized form. F1-score and accuracy are the base metrics for model evaluation for this experiment.

	precision	recall	f1-score	support
0	0.99	0.92	0.95	116372
1	0.12	0.52	0.19	2501
accuracy			0.91	118873
macro avg	0.55	0.72	0.57	118873
weighted avg	0.97	0.91	0.94	118873

F1 score: 0.1920529801324503
Precision: 0.11768419154116692
Recall: 0.5217912834866053
AUPRC: 0.12067391262045657
AUC: 0.8264378691516554
Accuracy: 0.9076325153735499

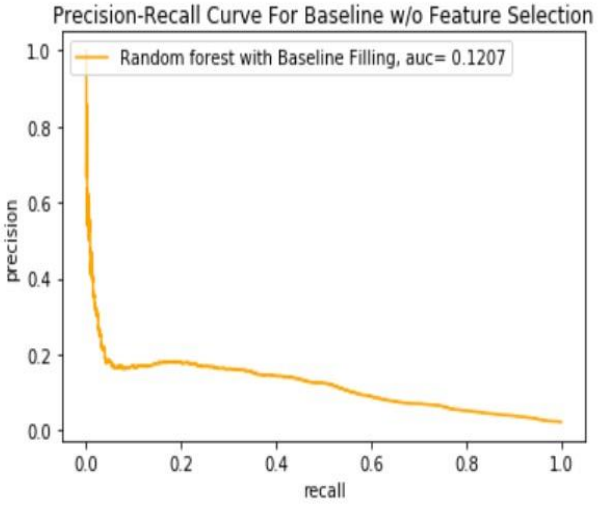

Fig. 14(a). Evaluation metrics for Model 1

In Fig. 14(a), the evaluation metrics for Model 1, which is the baseline RF model without feature selection and class balancing, are represented. The precision for false class is 0.99 and for true class is 0.12. The recall for false class is 0.92 and for true class is 0.52. For true label class, precision is low, and recall is moderate. This model with considerably low precision and moderate recall shows that it is returning moderate number of results but most of those results are incorrect predicted labels. These values of precision and recall are highly reflected in f1-score, as the value of f1-score is 0.19. Although the model has an accuracy of 0.9, it is rendered useless with very low value of f1-score.

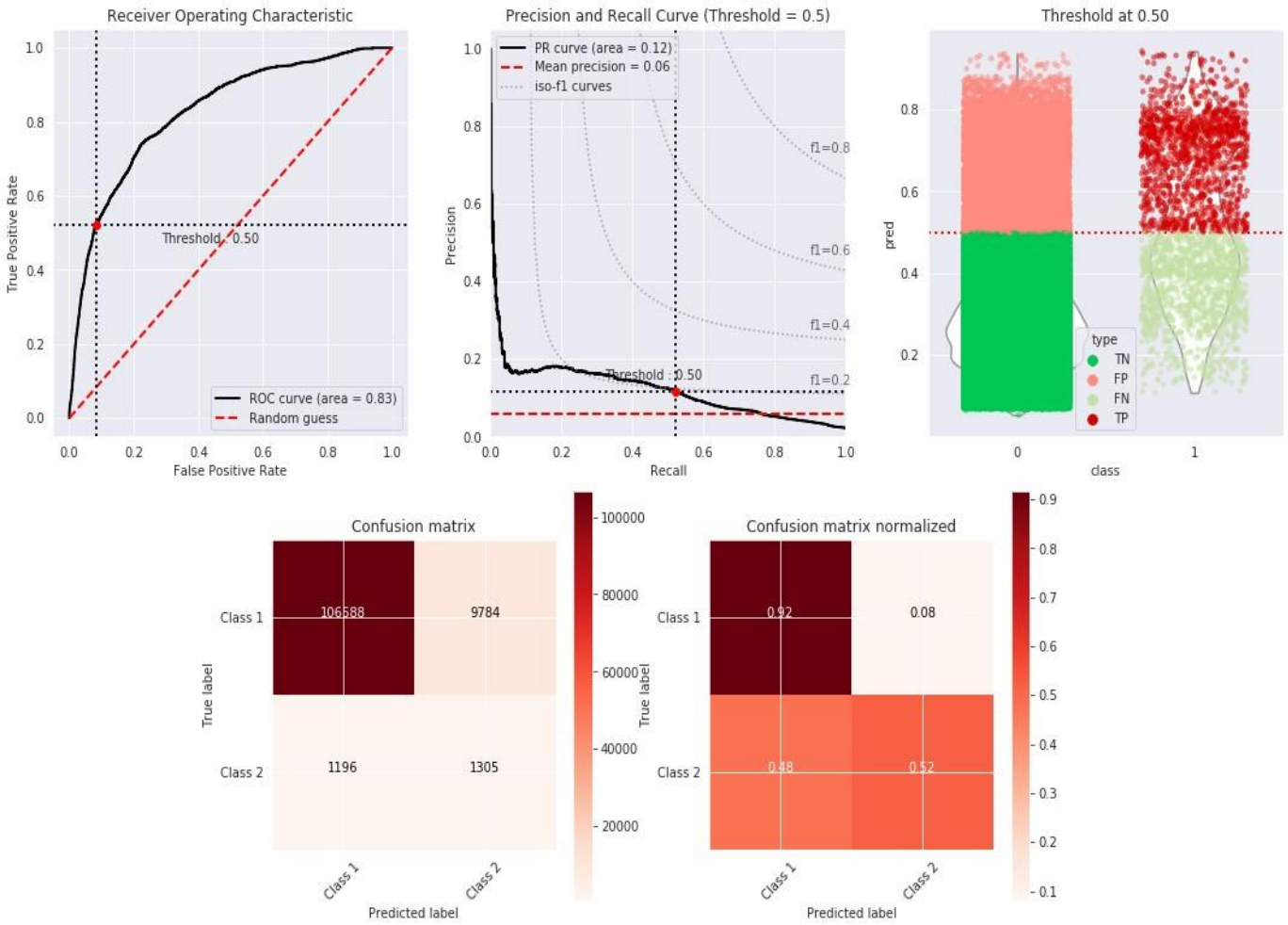

Fig. 14(b). Graphs of ROC curve, PR curve and Confusion Matrices for Model 1

The graphs in Fig. 14(b) are for POC curve, PR curve and Confusion matrices for Model 1. The value of area under curve for ROC curve (AUROC) is 0.83 and the value of area under curve for PR curve (AUPRC) is 0.12 with a mean precision of 0.06. The difference in area under curves for ROC curve and PR curve is due to the imbalance in the dataset, as ROC curve is better for more balanced dataset, and we are dealing with highly imbalanced dataset for Model 1.

	precision	recall	f1-score	support
0	0.99	0.92	0.95	116372
1	0.12	0.51	0.19	2501
accuracy			0.91	118873
macro avg	0.55	0.71	0.57	118873
weighted avg	0.97	0.91	0.93	118873

F1 score: 0.1888847994111152
Precision: 0.11575243594370263
Recall: 0.5129948020791684
AUPRC: 0.11899075921223842
AUC: 0.8210680186729833
Accuracy: 0.9073044341440024

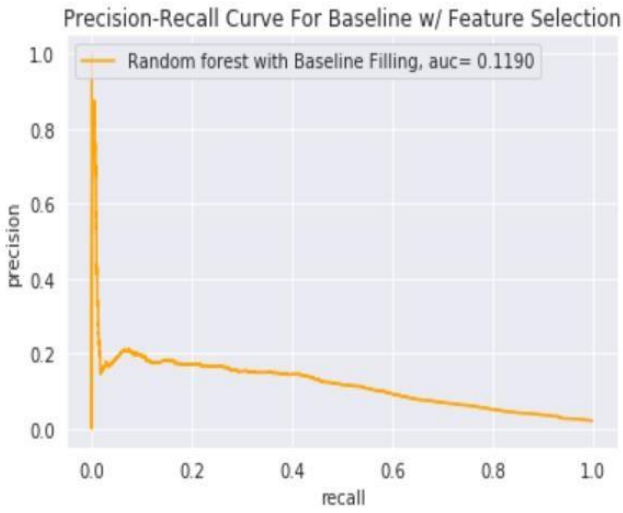

Fig. 15(a). Evaluation metrics for Model 2

In Fig. 15(a), the evaluation metrics for Model 2, which is the baseline RF model with feature selection and without class balancing, are represented. The precision for false class is 0.99 and for true class is 0.12. The recall for false class is 0.92 and for true class is 0.51. For true label class, precision is low, and recall is moderate. The values of precision and recall for Model 2 coincide with that of Model 1 which means that this is also a case of considerably low precision and moderate recall shows that it is returning moderate number of results but most of those results are incorrect predicted labels. These values of precision and recall are highly reflected in f1-score, as the value of f1-score is even reduced from 0.19 to 0.188. Although the model also has an accuracy of 0.9, it is rendered useless with an even low value of f1-score.

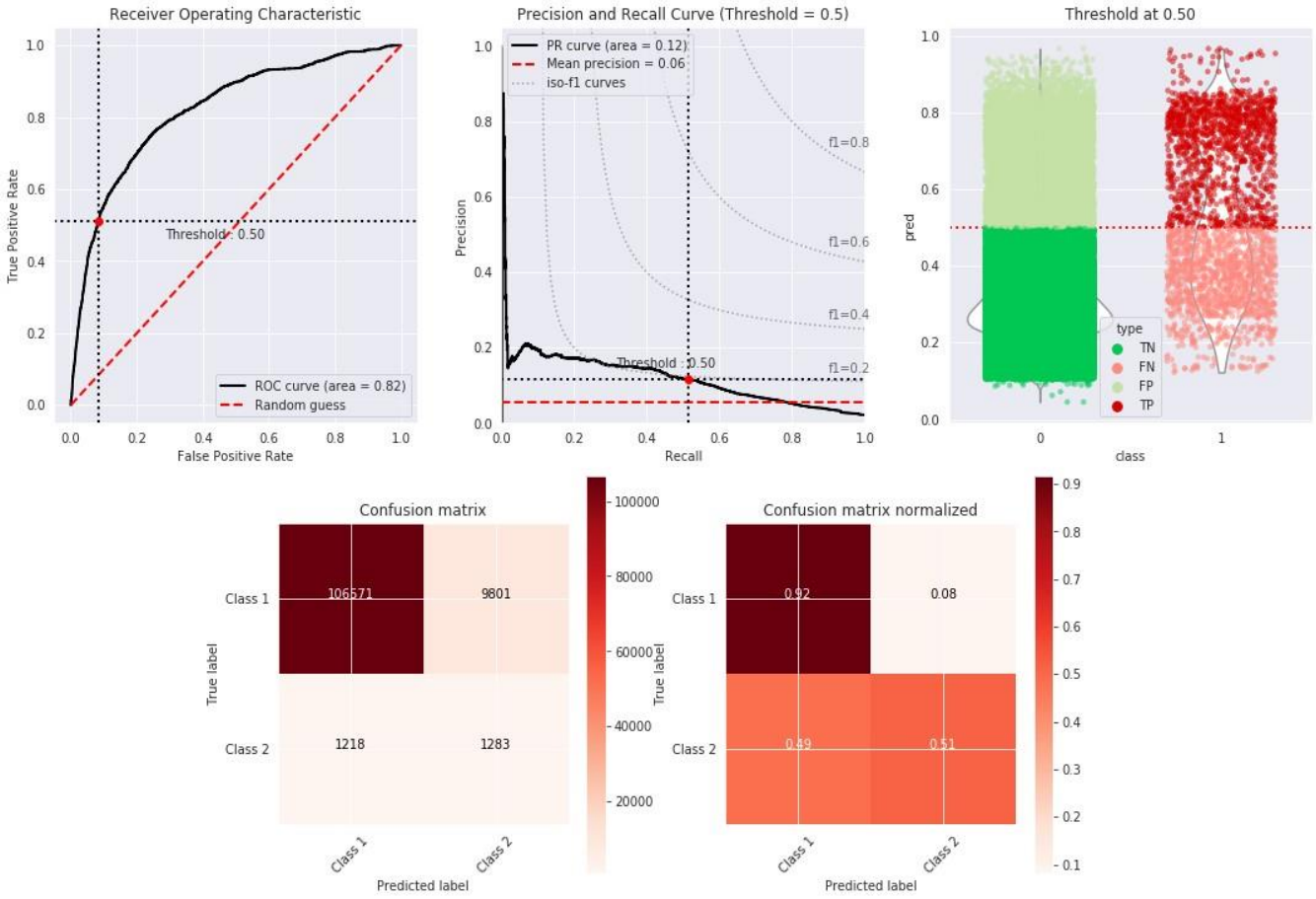

Fig. 15(b). Graphs of ROC curve, PR curve and Confusion Matrices for Model 2

The graphs in Fig. 15(b) are for POC curve, PR curve and Confusion matrices for Model 2. The value of area under curve for ROC curve (AUROC) is 0.82 and the value of area under curve for PR curve (AUPRC) is 0.121 with a mean precision of 0.06. Model 2 is also a case of imbalanced classes, therefore the value of area under ROC curve is not a valid evaluation metric.

	precision	recall	f1-score	support
0	0.65	0.91	0.76	114965
1	0.85	0.50	0.63	114965
accuracy			0.71	229930
macro avg	0.75	0.71	0.69	229930
weighted avg	0.75	0.71	0.69	229930

F1 score: 0.632315001172691
Precision: 0.8477491444116184
Recall: 0.5041882312008003
AUPRC: 0.8162291340450043
AUC: 0.8239110920226612
Accuracy: 0.7068194667942417

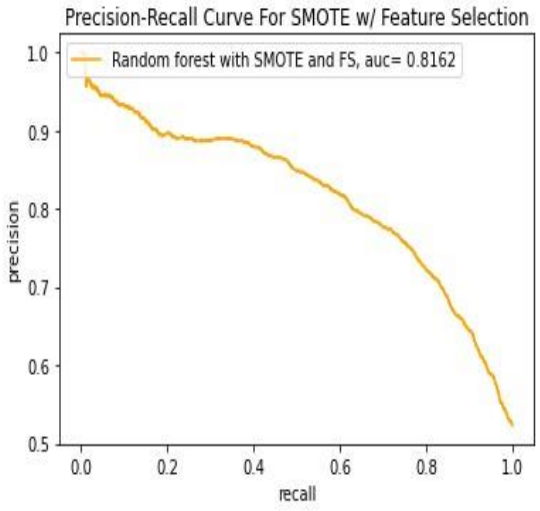

Fig. 16(a). Evaluation metrics for Model 3

In Fig. 16(a), the evaluation metrics for Model 3, which is the baseline RF model with feature selection and SMOTE class balancing, are represented. The precision for false class is 0.65 and for true class is 0.85. The recall for false class is 0.91 and for true class is 0.50. For true label class, precision is high, and recall is moderately high. This model has a very high precision value and moderately high recall value which means it nearly represents an ideal classification system where it will return large number of results with high percentage of predicted results labelled correctly. The F1-score for this model is 0.63 which is significantly better than F1-score of Model 1 and Model 2.

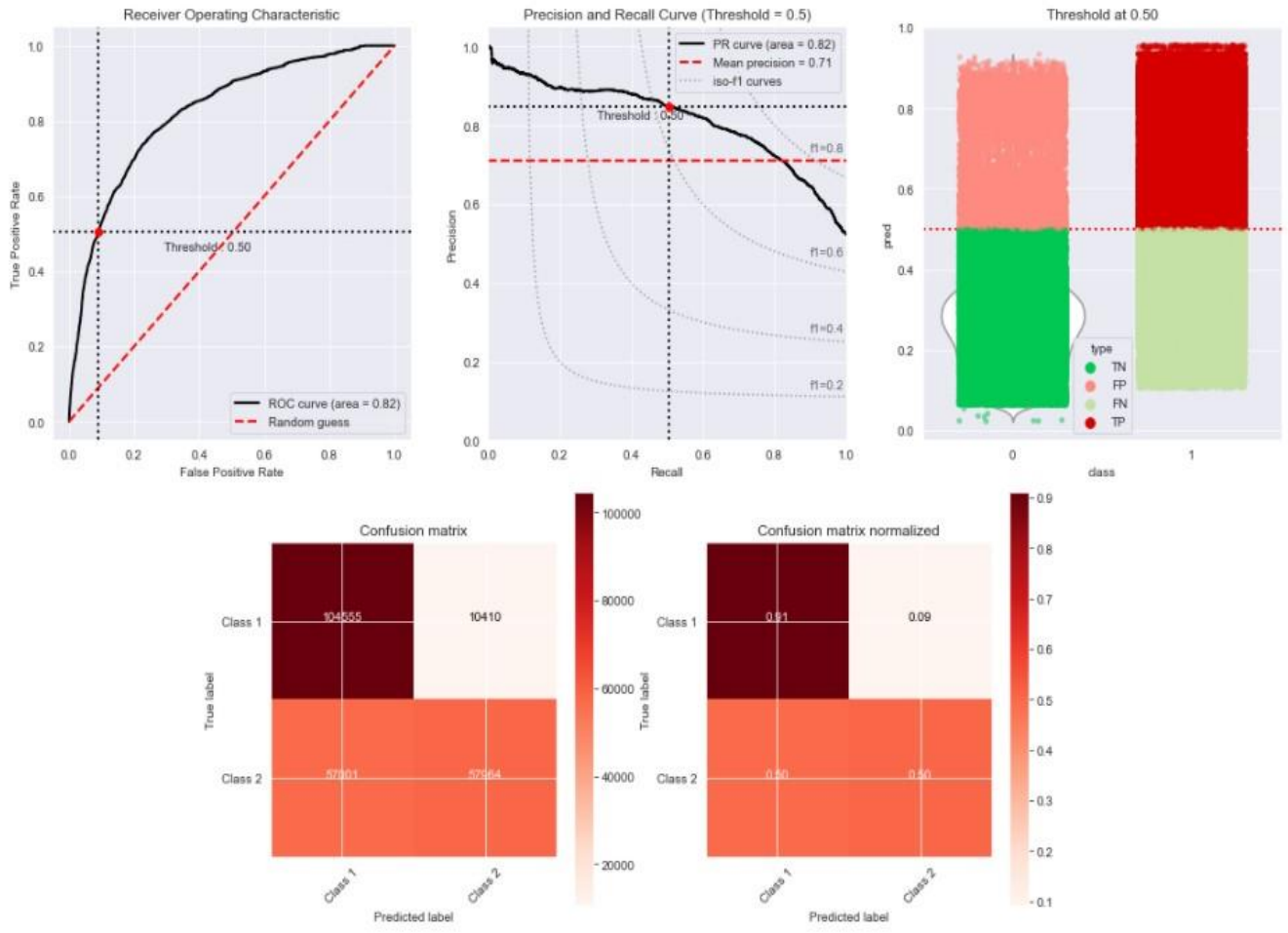

Fig. 16(b). Graphs of ROC curve, PR curve and Confusion Matrices for Model 3

The graphs in Fig. 16(b) are for POC curve, PR curve and Confusion matrices for Model 3. The value of AUROC is 0.82 and the value of area under PR curve is also 0.82. There is significantly negligible difference in the values of AUROC and AUPR. These values are important in the case of this model as the classes are balanced. (Result of SMOTE operation for Model 3)

Chapter 6: Conclusion

In our experiments of Geospatial analysis of hospitals in California, we can conclude that EDA methods are able to locate and identify medical facilities with high and growing SSI rates. The trends in SSI rates can be helpful in identifying these facilities and taking appropriate measures to reduce the SSI rates. One of the measures is to incorporate the use of early SSI detection systems as proposed in our study.

In our experiments, we saw that Model 1, which is the baseline RF model without feature selection and class balancing, and Model 2, which is the baseline RF model with feature selection and without class balancing, both had a very low precision value and moderate recall value. Both the models had very low value of F1-score despite their higher accuracies. Model 3, which is the baseline RF model with feature selection and SMOTE class balancing, showed high precision value and relatively high recall value. Model 3 outperformed both Model 1 and Model 2 in terms of F1-score.

The best model achieved an AUROC of 0.826 which is better than the values from the sepsis classification model in previous studies (0.64, 0.79, 0.81).[18,19,20] From this we can conclude that Random Forest turned out to be a better algorithm for sepsis labeling when combined with feature selection and class balancing. The f1 -score significantly increased (0.20 to 0.63) on performing class balancing using SMOTE, while accuracy decreased. The impact on accuracy can be explained by the size of dataset used as from a total dataset size of 40,000, only 20,000 samples were used in training the model due to computational limitations.

During this study, a hypothesis was made that the use of a diversified dataset, which includes vital signs of patients undergoing surgery, is the key to better sepsis labeling. It can be said that this hypothesis is proven true with this experiment. The hypothesis of obtaining a better time frame to predict the onset of sepsis failed due to constraints in the dataset such as the difference in recording frequency of Vital signs and laboratory values. The prediction model devised in this research performs better than the study referenced.[14]

This study can be concluded by stating that random forest as a tool for ensemble classification can be used with contiguous medical records consisting vital signs for identifying sepsis labels correctly, which ultimately solves the problem of early identification of surgical site infections.

6.1 Limitations and Future Work

In conclusion to our experiments, it was seen that the baseline random forest model was able to perform well when tools like feature selection and SMOTE class balancing were applied. Out of a total of 40,000 samples, only half the size of dataset was used. There is a great chance of a boost in performance if the complete dataset is used in model training, because the f1-score showed significant increase. In future, this model can be improved even further by using Advanced ensemble learning algorithms which are derived from self-learning ensemble method of boosting, which are GBM, Light GBM, XGBM, AdaBoost, and Cat Boost. With the study's methodology, models run concurrently and independently of one another, whereas in boosting, models run sequentially and rely on earlier models. Therefore, an experiment can be done to see how this impacts the learning rate.

The geospatial analysis done in this study had certain restrictions on the size of data used due to computational limitations. This geospatial analysis module can be refined by using the vast data from previous years which is available on the CDC database. This can significantly improve the trends identified in SSI rates of medical facilities spreading throughout the counties in California.

As discussed in previous sections, vital signs, which are a part of medical records, have a great co-relation with the early sepsis identification problem.[7] The availability of medical records stands in question, which has great impact on the learning rate on machine learning models in this study. Most of the medical records in hospitals are not readily available for research purposes due to security of data as the data is not anonymized.

Even if the data is available for research use, the data is inconsistent due to the difference in frequencies of how the data is recorded and stored. In recent studies, blockchain is being incorporated for managing and storing anonymized electronic health records, which provide security and greater accessibility of medical data.[25] Such tools can be combined with this study to yield better results and improve the standard of datasets used in machine learning systems. A new gold standard dataset or a data acquisition technique incorporating the above tool is required in future, without any constraints in the recording frequency differences of features.

References

[1] John Hopkins University Medicine. Health. Conditions and Diseases, Surgical Site Infections https://www.hopkinsmedicine.org/health/conditions-and-diseases/surgical-site-infections

[2] Sepsis Alliance. Sepsis and Surgery. 2022 retrieved from https://www.sepsis.org

[3] Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases (NCEZID), Division of Healthcare Quality Promotion (DHQP) 2010 retrieved from https://www.cdc.gov/hai/ssi/ssi.html

[4] Chopra, Teena & Jing, Jie & Alangaden, George & Wood, Michael & Kaye, Keith. (2010). Preventing surgical site infections after bariatric surgery: Value of perioperative antibiotic regimens. Expert review of pharmacoeconomics & outcomes research. 10. 317-28. 10.1586/erp.10.26.

[5] Sepsis - The American Journal of Medicine https://www.amjmed.com/article/S0002-9343(07)00556-6/fulltext

[6] Bone, Roger C., et al. "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis." Chest 101.6 (1992): 1644-1655.

[7] Hébert, A., Boucher, P., Guimont, C., & Weiss, M. (2017). Effect of measuring vital signs on recognition and treatment of septic children. Paediatrics & Child Health, 22(1), 13-16. https://doi.org/10.1093/pch/pxw003

[8] Berríos-Torres SI, Umscheid CA, Bratzler DW, et al. Centers for Disease Control and Prevention Guideline for the Prevention of Surgical Site Infection, 2017. JAMA Surg. 2017;152(8):784–791. doi:10.1001/jamasurg.2017.0904

[9] Urban JA. Cost analysis of surgical site infections. Surg Infect (Larchmt). 2006;7 Suppl 1:S19-22. doi: 10.1089/sur.2006.7.s1-19. PMID: 16834543.

[10] https://eloquesthealthcare.com/2018/07/11/financial-impact-of-surgical-site-infections-ssis/

[11] CDC Data Report Available at: https://www.cdc.gov/sepsis/datareports/index.html

[12] Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E. Epidemiology and Costs of Sepsis in the United States-An Analysis Based on Timing of Diagnosis and Severity Level. Crit Care Med. 2018 Dec;46(12):1889-1897. doi: 10.1097/CCM.0000000000003342.

[13] Division of Healthcare Quality Promotion, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention [Distributor: Healthcare-Associated Infections Program, California Department of Public Health]

[14] Early Prediction of Sepsis from Clinical Data -- the PhysioNet Computing in Cardiology Challenge 2019 https://physionet.org/content/challenge-2019/1.0.0/

[15] Hanselmann, Michael & Köthe, Ullrich & Kirchner, Marc & Renard, Bernhard & Amstalden van Hove, Erika & Glunde, Kristine & Heeren, Ron & Hamprecht, Fred & Morgan, Russell. (2009). Towards Digital Staining using Imaging Mass Spectrometry and Random Forests-Technical Report.

[16] Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):762–774. doi:10.1001/jama.2016.0288

[17] Vincent, J-L., et al. "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure." (1996): 707-710.

[18] Metsker, O., et al. “Sepsis Prediction Using Machine-Learning Methods: Prolonged Disorders of Consciousness Patients.” Journal of the Neurological Sciences, vol. 405, Elsevier B.V, 2019, pp. 83–83, doi:10.1016/j.jns.2019.10.1719.

[19] Goh, Kim Huat, et al. “Artificial Intelligence in Sepsis Early Prediction and Diagnosis Using Unstructured Data in Healthcare.” Nature Communications, vol. 12, no. 1, Nature Publishing Group, 2021, pp. 711–711, doi:10.1038/s41467-021-20910-4.

[20] Mao Q, Jay M, Hoffman JL, et alMulticentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICUBMJ Open 2018;8:e017833. doi: 10.1136/bmjopen-2017-017833.

[21] Re-examining causes of surgical site infections following elective surgery in the era of asepsis Prof John C Alverdy, MD, Prof Neil Hyman, MD, Prof Jack Gilbert, PhD. Published:January 29, 2020DOI:https://doi.org/10.1016/S1473-3099(19)30756-X

[22] Hébert A, Boucher MP, Guimont C, Weiss M. Effect of measuring vital signs on recognition and treatment of septic children. Paediatr Child Health. 2017 Mar;22(1):13-16. doi: 10.1093/pch/pxw003. Epub 2017 Mar 30. PMID: 29483789; PMCID: PMC5819837. [23] Kang H. The prevention and handling of the missing data. Korean J Anesthesiol. 2013 May;64(5):402-6. doi: 10.4097/kjae.2013.64.5.402. Epub 2013 May 24. PMID: 23741561; PMCID: PMC3668100.

[24] Chawla, Nitesh & Bowyer, Kevin & Hall, Lawrence & Kegelmeyer, W.. (2002). SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR). 16. 321-357. 10.1613/jair.953.

[25] Kshirsagar, Meghana & Patil, Aditya & Deshmukh, Saurabh & Vaidya, Gauri & Rahangdale, Mayur & Kulkarni, Chinmay & Kshirsagar, Vivek. (2020). Mutichain Enabled EHR Management System and Predictive Analytics. 10.1007/978-981-15-0077-0_19.

[26] University of Rochester Medical Center Data sources. SSI and effects

[27] Centers for Disease Control and Prevention Data Reports Available at: https://www.cdc.gov/sepsis/datareports/index.html

[28] Luo Ruisen et al 2018 IOP Conf. Ser.: Mater. Sci. Eng. 428 012004

[29] Sepsis: The evolution in definition, pathophysiology, and management available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6429642/

[30] Angus, Derek C., et al. "Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care." Critical care medicine 29.7 (2001): 13031310.

[31] Nemati, Shamim, et al. "An interpretable machine learning model for accurate prediction of sepsis in the ICU." Critical care medicine 46.4 (2018): 547.

[32] Islam, Md Mohaimenul, et al. "Prediction of sepsis patients using machine learning approach: a meta-analysis." Computer methods and programs in biomedicine 170 (2019): 1-9.

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

[34] Ibrahim, Zina M., et al. "On classifying sepsis heterogeneity in the ICU: insight using machine learning." Journal of the American Medical Informatics Association 27.3 (2020): 437443.