Predicting Surgical Site Infections Using Machine Learning Approaches with Further Investigation of Bias

By

VISHAL ILANGOVAN BHUVANESWARI

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Electrical and Computer Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Chen-Nee Chuah, Chair

Prabhu Shankar

Lifeng Lai

Committee in Charge

2021

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Predicting Surgical Site Infections Using Machine Learning Approaches with Further Investigation of Bias

Digitalization of healthcare records has made patient-centered records, commonly known as Electronic Health Records (EHRs), readily available and has provided opportunities for secondary analysis. These EHRs provide us with patient demographics, laboratory values, patient vitals values, the medications administered throughout the treatment, the type of surgery the patient underwent, outcomes, and much more information. This readily available data has enabled the fast-growing field of Machine Learning and Artificial Intelligence to build reliable patient statistics and gain useful insights which support healthcare providers in making better decisions, thereby improving the quality of healthcare.

This Thesis demonstrates one such use case of EHR data in predicting the onset of Surgical Site Infections (SSIs). The objective is to predict and stratify patients who are at risk of developing SSI by applying various Machine Learning (ML) methods.

Surgical Site Infections (SSIs) can be defined as the infections that occur at the site of the surgery within 30 to 90 days of the procedure, depending on the type of the procedure. SSIs account for about 20% of all Hospital-Acquired Infections (HAIs) and have an enormous effect on patients, hospitals, and public health in general. Predicting who may be at risk for SSIs can help clinicians take preventive measures to avoid the onset of the infection. Availability of data at both the pre-and post-operative stages allows the application of ML methods at each of the stages, aiding in drawing better insights.

The usage of this readily available data comes with its own downsides. There is an inherent bias in data that could have been induced at different stages of the data acquisition process. These biases, when left unaddressed, can creep into the algorithms we employ and result in biased decisions. This unintentional bias may seem unfair to certain groups of the patient population. Identifying and mitigating this bias issue will result in a “fair” predictive model. In this Thesis, we analyze and demonstrate the issue of ML bias in using retrospective patient data.

ACKNOWLEDGMENTS

I would like to thank Prof. Chen-Nee Chuah for allowing me to work on this research project. Thanks for being patient with me, guiding me along the way with insightful comments, and helping me see the clear picture during times of uncertainty.

I extend my sincere thanks to Dr. Prabhu Shankar for having a great deal of belief in me and always being available to answer my questions. I enjoyed our meetings with a good balance of casual and technical talks, and I would always leave with a boost of confidence and a sense of optimism. Thanks again for always supporting me with your kind words and for all your help.

Thanks to Prof. Lifeng Lai for agreeing to serve on this committee. Your course on Reinforcement Learning was one of the most interesting and challenging courses I have taken. It helped me explore another subset of Machine Learning and further build my skillset.

Thanks to Dr. Debraj Basu and Dr. Chao Wang for all their time and valuable comments and suggestions. Thanks for clearing my doubts and pointing me in the right direction to develop my skills in the vast field of data science.

CHAPTER 1: INTRODUCTION

1.1 Problem Statement

Surgical Site Infections (SSIs), defined as the infections that occur at the site of the surgery within 30 to 90 days of the procedure can have an enormous negative impact on patients, hospitals, and public health in general. Machine learning approaches, and further analysis of inherent biases may help with predicting who may be at risk of SSI and help clinicians take necessary interventions to prevent the onset of SSI.

1.2 Thesis Statement

The objective of this thesis is to predict and stratify patients who are at risk of developing SSI by identifying the risk factors and applying various Machine Learning (ML) methods. Further, we also analyze and demonstrate the issue of ML bias in using retrospective patient data fpr the task of SSI classification.

1.3 Approach

The goal of the experiments performed in this thesis are to provide a model for SSI risk prediction and to handle the inherent bias present in data modeling. Our approach begins with reducing data from two patient data sources, the SSI surveillance National Healthcare Safety Network (NHSN) registry data, and the associated Electronic Health Records (EHR), into a single covariate matrix. The covariate matrix was further processed, and data inconsistencies were addressed. Next in the pipeline was a two-step feature selection process that filtered the important features based on statistical measurements and further reduces the feature set by factoring in the

downstream classification task. This process provided us with a set of predictors that contribute the most towards SSI risk prediction. The final step of the pipeline involved the Nested cross-validation process that provides a robust estimate of the performance of the ML model. All the input features for the model were structured and discrete providing an opportunity to automate the process of prediction to provide decision support in real-time at the point-of-care across the continuum of pre-operative surgical planning phase through to the post-operative recovery phase.

In the above-mentioned experiments, we do not deal with the inherent bias present in the data. The next part of this thesis deals with identifying and mitigating this bias. Identifying this bias begins with defining the subgroups from the patient population followed by assessing the ML model's performance across these groups to provide proof of the existence of bias in terms of statistical metrics. Further, we discuss mitigation techniques to handle the bias issue.

The approaches in this thesis were carefully designed in a way that makes sense in both, engineering as well as in the medical field, by working closely with clinical subject experts.

1.4 Organization

This thesis is organized into 6 chapters. Chapter 2 details the technical methodologies and algorithms used in our experimentation. In Chapter 3, we discuss the background, significance, and related works in SSI prediction and the algorithmic bias issues. Chapter 4 details the processes involved in developing the pipeline for SSI prediction along with the results. Chapter 5 explains the methods employed in bias identification and mitigation along with the results. Chapter 6 concludes and discusses the interpretation of the results and the limitation in the experiments from the previous chapters.

CHAPTER 2: TECHNICAL CONCEPTS

2.1 Predictive Modeling

2.1.1 Random Forest (RF)

A decision tree, as the name suggests, models the data in the shape of a tree which is then used to classify a new unknown data point. For a binary classification tree, at every node, the input training data points are split into homogenous groups based on the features in the dataset. The "Gini" score, similar to a cost function, indicates the quality of a split. The Decision tree has a recursive nature and is a greedy algorithm with the objective to reduce the cost of splitting. With such an algorithm there rises a possibility of overfitting, where the algorithm perfectly models the given data but fails at generalizing.

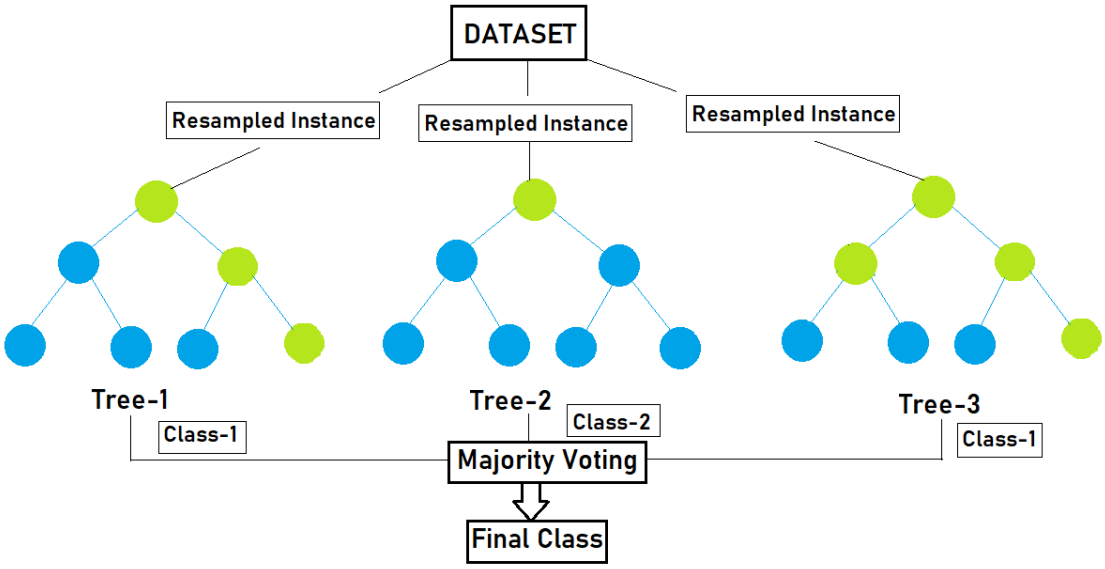

Figure 1. Random Forest

The Random Forest (RF) algorithm encompasses multiple decision trees. An RF model consists of a cluster of decision trees and the final prediction, either classification or regression, is based on all the trees. Each decision tree in RF is built using a subset of the input training data points obtained by resampling and a random set of features (See Figure 1). The output of an RF classifier is the class with the majority of the votes from the cluster of trees. By using a random vector of features for each tree RF decreases the correlation between the trees which in turn reduces the chances of overfitting and helps in generalizing the model to unknown data points better than a decision tree [1]. But there is still the chance of individual trees overfitting that can be handled by manipulating the characteristics of the trees such as the depth and the minimum number of data points per leaf node.

2.1.2 Convolutional Neural Network (CNN)

Artificial Neural Networks (ANNs) are a category of algorithms that use an interconnected structure of neurons to learn patterns from the data. ANNs are often compared to the neural connections in a human brain and hence the name. The structure often consists of an input layer, followed by multiple hidden layers, and then an output layer. Each layer of the network consists of a weight matrix (also known as parameters of the network) and reveals features of interest, from the input data, that help in deducing the patterns and aids in the task of classification. The input data is transformed from one layer to another often by a linear matrix multiplication with the weight matrix followed by a non-linear activation. The activation function is basically a thresholding function that activates a neuron and allows data transfer from that neuron if the value in the neuron is above a specified threshold. The output layer produces the final result, and for a classification task, the size of the output layer depends on the number of classes in our data.

$$cross - entopy\ loss = \frac{-1}{output\ size} \sum_{i=1}^{output\ size} y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)$$

The network learns the patterns by a training regime that consists of two processes: forward propagation and backward propagation. The forward propagation produces the final result by propagating the data through the layers. The backward propagation involves a cost function that calculates the cost of misclassification. The weight matrices of the layers in the network are then updated as a function of this cost. The network is repeatedly trained with these two processes with the objective to minimize the cost of misclassification. We employed a cross-entropy loss function as defined by above equation. In the equation, y_i denotes the true class and $\hat{y}_i$ denotes the class predicted by our network for the given input sample. The training regime also involves several hyperparameters that help in manipulating the quality and duration of training.

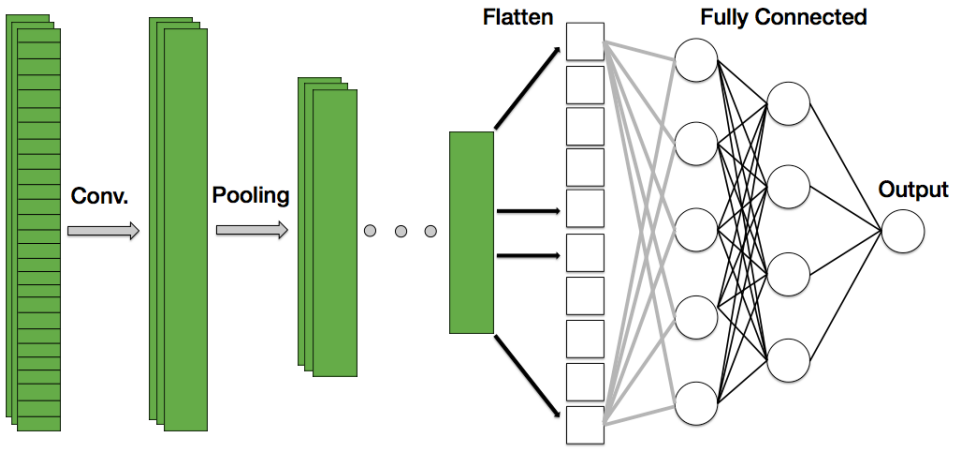

Figure 2. One-dimensional CNN

Convolutional Neural Networks (CNNs) are a subset of ANNs that performs a convolution operation between matrices in the forward propagation of the network. At every layer, the input to

the layer is convolved with the weight/parameter matrix of the layer. A simple CNN, as shown in Figure 2, consists of multiple layers: convolutional layer, pooling layer, non-linearity layer, and fully-connected layer [2]. Fully-connected layers are the same as the neuron arrangement found in a traditional neural network. The convolutional layers and the fully-connected layers are the ones containing the parameters which are updated during the training of the network. The training of a CNN is quite similar to that of an ANN with the forward propagation and the backward propagation.

CNNs with multiple deep layers are known for their usage in the classification of unstructured data such as image and text data but in this thesis, we have implemented a shallow one-dimensional CNN for the task of classifying structured data.

2.2 Feature Selection

2.2.1 Statistical Filtering

The Kolmogorov-Smirnov (KS) test is used to test for the goodness of fit of one sample distribution with another. Given a sample and a reference or two samples, the KS test for goodness of fit tests the null hypothesis: *the two samples were drawn from the same distribution*. The motivation to test the goodness of fit is supported by the argument that to classify the records into positive and negative classes, we need features that reflect the differences in distribution and help the model distinguish between the two classes. The main advantage of the KS test is that it is a non-parametric and distribution-free test, that is, it is independent of the underlying distribution of the data [3].

In the KS test, we divide our feature samples into two groups based on the target to create empirical Distribution Functions (EDFs). The vertical distance between these two distributions (d)

is calculated, as observed in Figure 3 (Bscan, 2013) [40]. This distance (d) is then compared with a critical value that can be obtained readily from the KS-test P-value table using the sample size and the significance level. If the distance (d) is greater than the critical value, then the hypothesis can be rejected. The features that reject the null hypothesis can be considered to distinguish between the binary target values.

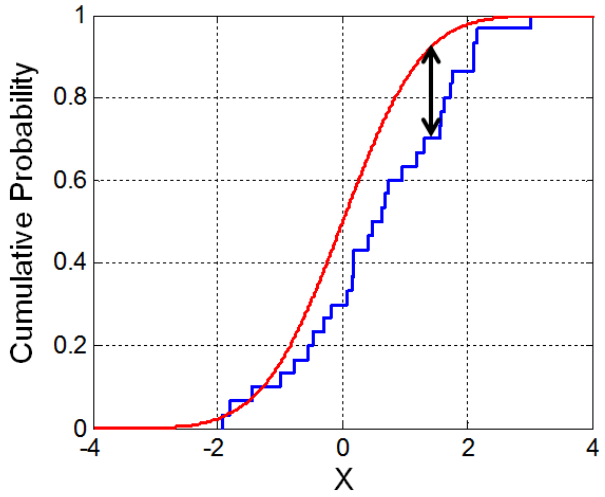

Figure 3. KS statistic depicting the vertical distance between two distributions (Bscan, March 2013)

The KS test is a univariate test that helps in proving the existence of a correlation between the features and the target variable. Although this helps in choosing the relevant features, there might exist a correlation between some of the chosen features which may not contribute to the prediction of the target variable. These highly correlated features need to be excluded from the analysis using multivariate test, such as Pearson's correlation test.

The Pearson's product-moment correlation coefficient is a numerical value indicating the linear association/correlation between two features. The values range from +1 to -1. A positive value indicates a positive relationship, that is, as one variable increases, the other also increases, and similarly, a negative value indicates a negative relationship, that is, as one variable's value

increases, the other decreases. A higher absolute value indicates a high correlation between the two features. A value close to zero indicates no correlation between the two variables [4].

The correlation coefficient is calculated based on the spread, that is, the variation of the data points, around the line of best fit drawn using the two variables. The three possible cases of correlation are depicted in Figure 4 (Steven Nickolas, Feb 2019) [41]. A higher spread leads to a lower correlation coefficient value and a lower spread, that is, the data points showing low variation away from the line, lead to a higher absolute coefficient value. Based on the obtained values, an empirical threshold is set to filter out the features that are least correlated with one another.

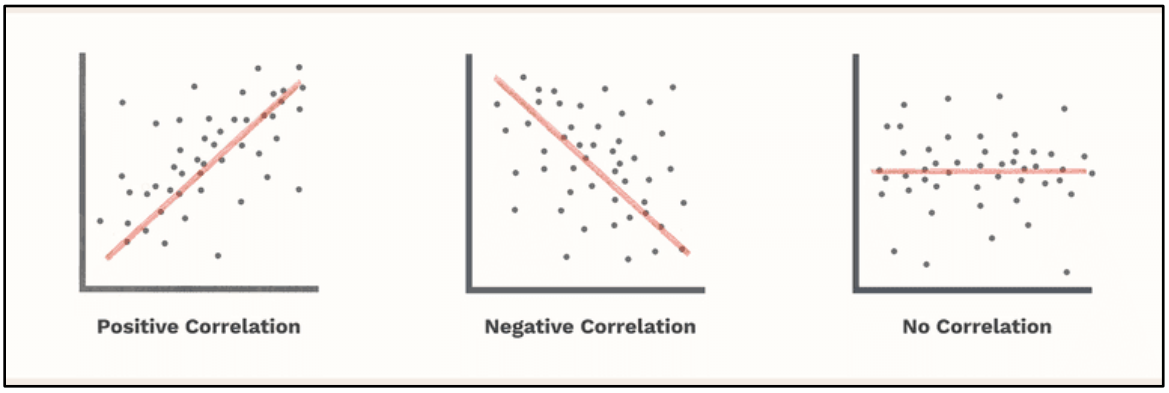

Figure 4. Correlation depicted by data spread and the line of best fit (Steven Nickolas, Feb 2019)

2.2.2 Maximum Relevance and Minimum Redundancy (MRMR)

Maximum Relevance and Minimum Redundancy (MRMR) is a statistical feature selection method that picks up features based on their relevancy or correlation with the target variable and avoids redundant features [5]. Redundant features are those that have high cross-correlation with other features and do not provide any new information but might add noise to predictive modeling. This is done in an iterative manner, where the number of iterations depends on the number of features required. There are two steps at every iteration, first, the correlation between each feature

and the target is calculated (c1). Second, the cross-correlation between the features that have not been selected yet and the already selected features is calculated (c2). The final score is calculated as c1/c2, and the feature with the highest final score is selected at that iteration.

Following the MRMR feature selection, we implemented an exhaustive feature search method known as forward selection. This is an exhaustive search method that iteratively selects features that when combined with already selected features improve the predictive performance of the model. Unlike the previous selection algorithm, this method takes into consideration the downstream classification task. This is a suitable method when the number of features is limited as the computational complexity is high. By using the MRMR method, we were able to filter out a significant number of features and reduce the size of the feature set enabling the use of a more robust feature selection method such as forward selection, which otherwise would be computationally very expensive.

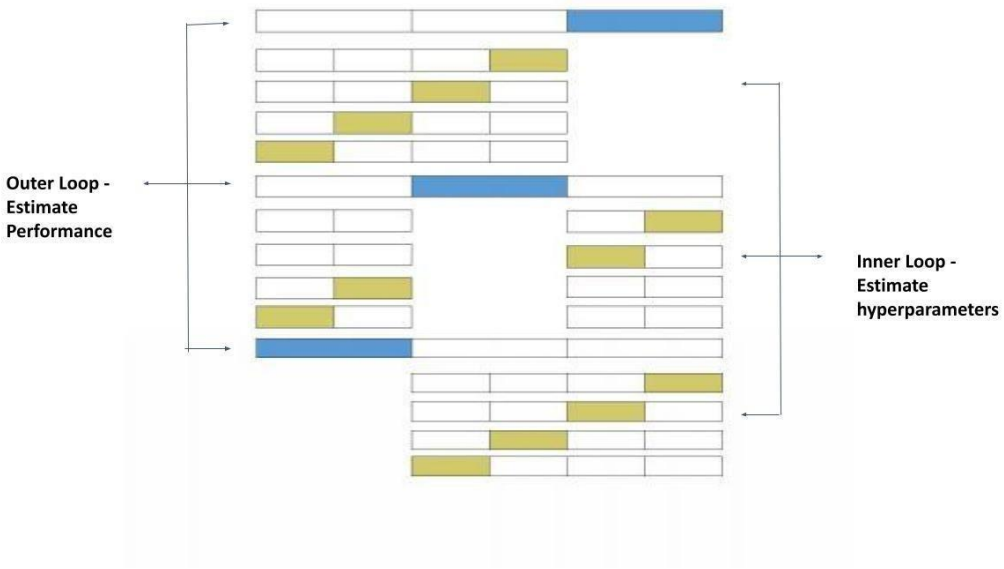

Figure 5. Nested Cross-Validation

2.3 Nested Cross-Validation

Nested Cross-Validation (NCV) was employed to provide a robust estimate of the generalized model performance. In NCV, there are two nested loops, outer and inner, the outer loop splits the data into ‘K’ folds and at each iteration one of the folds acts as a hold-out data and the remaining data is used for training. In the inner loop, the training data is split into ‘n’ folds and at each iteration one of the folds acts as validation data and the remaining folds are used for training. The purpose of the inner loop is to search for the best hyperparameter that improves the model’s performance on unseen data. The model is then fitted with the best hyperparameter found, trained on the outer loop’s training data, and evaluated using the outer loop’s hold-out data. Figure 5 shows a pictorial representation of the process.

CHAPTER 3: BACKGROUND AND RELATED WORK

3.1 Surgical Site Infections

Surgical Site Infections (SSI) are an adverse event that complicates 2% to 5% of patients who undergo inpatient surgery in the United States. Approximately 21.8 million surgeries are conducted, and around 160,000–300,000 SSIs occur annually in the United States [6, 7]. SSI is now the most common and most expensive Hospital-Acquired Infection (HAI), accounting for 20% of all HAIs in hospitalized patients.

The impact of SSI on patients, hospitals, and public health in general, is enormous. SSI significantly affects patient outcomes. Along with discomfort, pain, and suffering due to additional interventions, SSIs are estimated to add an extra 7 to 11 days to a patient's hospital stay, 2 to 11 folds higher risk of death compared to non-SSI patients, and 77% of deaths in patients with an SSI are directly linked to the SSI. Further, many of the costs associated with managing SSI are non-reimbursable. SSIs add about $3.5 billion to $10 billion annually to healthcare expenditures [6]. Therefore, SSIs are a significant public health concern due to its association with morbidity, mortality, and healthcare costs.

The quality improvement initiatives by public health institutions have generated evidence-based guidelines that provide strategies to prevent SSIs. It is estimated that up to 60% of SSIs are preventable by following these guidelines. Recent reports suggest a slight decrease in the incidence of SSI and other HAIs, after the implementation of practices that are following those guidelines, though more research and efforts are needed [8]. SSIs are a national healthcare priority, and several initiatives for SSI surveillance within hospitals have arisen over the past few decades, with the goal of improving the early detection of SSI [9].

Much of the research concerning SSI is in developing surveillance methods for monitoring SSI rates in hospitals. Early efforts at the monitoring of SSI by the CDC have helped develop standard metrics for tracking SSI rates and other surgical outcomes over time and across institutions so that trends and comparison of infection rates between facilities are standardized. Utilizing these standards, the CDC has established the National Healthcare Safety Network (NHSN) and developed standards to track HAIs, including SSI. The information system, National Nosocomial Infections Surveillance System (NNIS), was set up by CDC in 1992 and helps participating institutions follow standard definitions and guidelines at data collection, which is mandatory in most states and tied to payment determinations.

The American College of Surgeons followed the NHSN methods, pooled collective knowledge of its members to add additional data elements to the standard matrices and implemented the National Surgical Quality Improvement Program (NSQIP). The data capture methods at point-of-care for both NHSN and NSQIP are active surveillance and retrospective review of clinical documentation. Though the NSQIP surveillance methods differ slightly from the NHSN, both standards, definitions, and surveillance data are broadly accepted benchmarks, implemented across healthcare facilities as part of quality improvement initiatives and surgical outcomes research [10, 11]. Both data repositories provide a rich source of manually curated and validated data, capturing both SSI and non-SSI control populations from healthcare facilities across the US for research purposes.

The availability of validated data on SSI patients and the non-SSI control population has led to a range of efforts at developing generalizable forecasting and predictive modeling methods for identifying patients who are at increased risk for SSI. Typically, however, these efforts are limited to data captured using these NHSN and NSQIP databases. In addition, very few studies have

attempted to incorporate additional features into their modeling approaches, such as laboratory, vitals, and medication data [9, 12, 13, 14, 15].

Identification of risk factors is an essential step in the early detection of SSI. Previously known risk factors are based on patient- and surgery-related variables, with a few studies incorporating additional features such as laboratory, vitals, and medication data. Laboratory data is one of the most reliable indicators of health status, a measure of milieu interior, that clinicians have depended heavily on for decision-making about all aspects of patient management. Moreover, many studies focus on SSI occurrence in specific procedures and hence report surgery-specific risk factors. This project attempts to overcome some of the shortcomings of previous efforts both by incorporating additional data sources (from Electronic Health Records) and using modern analytic methods for a variety of surgical procedures. Combining and analyzing data from disparate patient health measures has thus yielded higher accuracy than similar methods that have not incorporated all available additional patient care data.

Artificial Neural Networks (ANNs) have proven to model complex non-linear patterns that linear models struggle to identify. ANNs are typically used for unstructured data such as images and text. Image analysis of wounds with ANNs for the detection of SSI has shown very good performance in [16, 17]. Though ANNs are preferred for unstructured data, efforts have been made to avail the ability of ANNs to model complex relationships in structured data analysis [8, 18, 19]. In this project, we attempt to use Convolutional Neural Networks (CNNs) [20], which are a subset of ANNs, in the prediction of SSI. Various other ML methods were studied, which are further discussed in the methods section.

3.2 Bias in ML

We also demonstrate the bias issue in ML as a part of our study. Algorithmic bias in ML, also popularly known as fairness in ML, is a currently emerging field of Artificial Intelligence that deals with the biases in data and model, that causes them to perform unevenly across groups/individuals that are defined based on the protected attributes such as race/sex. In a broad sense, the bias issue in ML is defined as the disparity in the model's performance, in terms of some statistical measure, among different groups/individuals [21]. This bias in the model performance can affect the decision-making process and hence result in certain groups being at disadvantage in terms of receiving medical care or attention. Such bias, when not diagnosed and left untreated, propagates further into future clinical decision-making processes that can negatively impact patient outcomes in certain group of patients.

There are many formal definitions of fairness in ML presented in the literature. The most common definitions are Demographic parity, Equality of Odds, and Equality of Opportunity [22, 23]. Each of these definitions deals with a metric, a measure of bias, bounded by a certain statistical constraint. A number of these metrics, depending on the context and the subject matter expertise, can be defined as a function of the confusion matrix [24]. The disparate impact, which is defined as the disproportionate impact of the outcomes on the subgroups [25], is calculated to quantify the disparity in the model performance. This disparate impact helps in identifying the existence of bias.

Once the existence bias has been identified, the next step would be to mitigate the said bias. Mitigating bias deals with satisfying the definition of fairness, that is, the statistical constraints that bound the metric of choice. Based on the type of constraints violated, the appropriate mitigation technique can be chosen. Mitigation techniques are usually grouped into pre-processing, in-

processing, and post-processing algorithms [26, 27]. Each algorithm manipulates a different part of the classification pipeline. In this thesis we demonstrate the identification and mitigation of bias in our SSI data.

CHAPTER 4: SSI RISK PREDICTION

4.1 Setup

In this chapter, we discuss the workflow and techniques employed in the task of SSI risk prediction. We first discuss the data collection and the data preprocessing which provides a description of the dataset and the decisions taken in processing the data. Then, we move onto the features selection section that explains the need for such a process in the pipeline and the analytic methods implemented in this thesis. The model development section describes the training of the classical ML algorithms and the CNN. Following that, the results section discusses the features obtained and the model performances.

4.2 Data Collection

After appropriate IRB approval and following standard operating procedures, the validated NHSN surveillance data of SSI patients and the denominator control non-SSI population was collected for all patients at the UC Davis Health affiliated hospitals who underwent surgery between 2014 and 2017. In addition to the NHSN SSI surveillance data, additional data was collected for each patient from the internal UC Davis Electronic Health Record (EHR) Clarity database, which includes laboratory, vitals, problems list, and medications data.

4.3 Data Preprocessing

The raw dataset contained separate files and representations (cohorts) for each patient representing disparate sources for laboratory, vitals, medication, and surgery related information

spanning from 2014 to 2017. It was necessary to clean and merge these data files. The data cleaning process was primarily performed using the R programming languages. Packages such as "tidyverse", "dplyr", and "tidyr" were used to perform appropriate cleaning of the data and handle any inconsistencies that may be present across the data from different years [28], such as the variation in data formats and inconsistencies in patient demographics.

Each patient had multiple values for labs and vitals that were taken at different time points for each procedure. Hence, to model these values in a time agnostic fashion, it was necessary to create derived variables that summarized these values for each procedure. The lab and vitals' values were represented by the minimum, maximum, mean, and the range of the lab and vitals measurements. So, each type of lab and vitals' measurement was represented by four derived features. Additionally, the differences in values of temperature and pulse, the two types of vitals variables, before and after the surgical procedure were calculated for every patient and included in the features list.

After generating the features, the categorical variables, such as the surgery-related and medication variables, were one-hot encoded to create dummy variables. The data sources with the newly generated variables were now combined into one covariate matrix with the "patient id" and "surgery date". In the reduced covariate matrix, each row represents an encounter and not a patient since each patient can have multiple procedures. Therefore, each row has a unique combination of "patient id" and "surgery date".

The current covariate matrix contained data of patients spanning from 30 days prior to their surgery to 7 days after the procedure. In the resulting covariate matrix, features with more than 40% of their values missing were removed, that is, measurements that were present for at least 60% of the procedures were retained. Our exclusion criteria also included removing patients under

the age of 18 years. The covariate matrix contained 32 different laboratory types that satisfied the retention criteria. The missing laboratory values were then imputed. The imputation technique was chosen contingent on the computational time complexity required to impute the values. In this study, data imputation was performed using the "Simple imputer" package in python [29]. The missing values in every column would be replaced with the median of that column. One main drawback of this method is that it reduces the variance of the data which could hinder learning, but further experiments with Random Forest-based and K Nearest Neighbours (KNN) based imputation algorithms showed similar performance in the downstream classification task. Table 1 shows an example representation of the categorical and the numerical features in the covariate matrix with every row representing one encounter.

Table 1. Representation of the patients' covariate matrix. "PID" stands for the patient ID and "NA" indicates missing values.

PID	L1	…	M1	…	V1	…	C1	…	S1	…	SSI (Class label)
1	23	…	0	…	90	…	1	…	0	…	1
2	NA	…	1	…	98	…	0	…	0	…	0
3	45	…	0	…	91	…	0	…	1	…	0
-	NA	…	0	…	98	…	1	…	0	…	1
-	20	…	1	…	96	…	1	…	0	…	0
n	NA	…	1	…	80	…	0	…	1	…	0

4.4 Feature Selection

After the creation of the cohort data, there were 290 feature variables. These features included numerical as well as categorical features. The categorical features were one hot encoded and dummy variables were created. Feature selection was necessary for the following reasons: 1) To reduce the feature space and thereby reduce the computational complexity 2) select the features that provide the most information by reducing the noise as an irrelevant feature affects the learning process and a redundant feature adds nothing new to the task of prediction. 3) Identify relevant features for the clinicians.

In this study, we tested and compared two data-driven features selection methods. The first one was a statistical two-step filter method. This is a combination of a univariate filter, that filters the relevant features based on their relationship with the target, and followed by a multivariate filter, that filters redundant features. We were able to reduce the size of the feature set that enabled model selection, which otherwise was computationally expensive. This is discussed further in the results section. Following that, we implemented a more robust combination of the MRMR technique followed by the exhaustive forward selection process. This combination, unlike the statistical method, takes into consideration the downstream classification task. The details on the implementation of these techniques can be found in Chapter 2.

4.5 Model Development

We tested four different classical ML classification algorithms such as Random Forest (RF), XgBoost, Logistic Regression, and AdaBoost. Additionally, we also tested the performance of Convolutional Neural Networks on the task of SSI classification.

The performances of the models were assessed using 5-fold Nested Cross-Validation (NCV). In each run, the data is split into stratified folds based on the patient id and the target variable such that each fold contains the same proportions of the positive and negative cases, and there is no overlap in patients between the folds**.**

Due to the high feature dimensionality, it was necessary to reduce the feature set to select the model. Using the features from the statistical method, we evaluated the performance of the models by the Area Under the Receiver Operating Characteristic (AUROC) curve, the Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity, and Specificity. Based on these metrics, we chose RF for further experimentation.

All the classical ML algorithms were implemented using their respective sci-kit learn packages to model the data for the task of SSI risk prediction. A NCV process, which encompasses 5-fold cross validation hyperparameter search, was employed to evaluate and compare the performance of the model across the two sets of features selected. By using NCV we can provide a robust estimate of the model performance. Chapter 2 describes the working of NCV in detail. The following section explains the techniques and tools used in the training of the CNN.

Convolutional Neural Network (CNN)

The PyTorch library was used to define the neural network model class and the training process [30]. The network architecture contains a 1D convolution layer, a fully connected layer, and a batch normalization layer. The Rectified Linear Unit or ReLU activation function was used as the non-linear activation of the network. The weights of the network were not initialized randomly and instead, they were initialized using a semi-orthogonal matrix that allows depth independent learning times and faithful propagation of gradients [31]. The size of the input was

the size of the entire features set, that is, 290 features. The input to the network was scaled since it helps in training the network efficiently.

$$weighted\ cross-entopy\ loss = \frac{-1}{output\ size} \sum_{i=1}^{output\ size} \alpha y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i)$$

SSI is a rare event, and our data has an SSI prevalence rate of 3.3%. To deal with this imbalance in class population between non-SSI and SSI cases, we used a weighted variant of the cross-entropy loss function. Adding a cost or a weight to a term in the cross-entropy loss that corresponds to the misclassification of the minority class penalizes the misclassification of a minority class with more scrutiny than that of a majority class, and hence enables the network to learn the minority class data well. The above equation represents the weighted cross-entropy equation with the weight term α that penalizes the misclassification of the minority class.

Hyperparameter tuning is an important part of neural network training. Depending on the data, a different configuration of hyperparameters may be required to achieve the best performance possible. Searching for the best set of hyperparameters can be exhausting and computationally expensive. In our implementation, we used an automatic hyperparameter tuning framework known as Optuna [32] that utilizes parallel processes or threads to search hyperparameter spaces and rules out discouraging configurations. Refer Chapter 2 for a brief explanation of the working of a neural network.

4.6 Results

The baseline characteristics and the demographics of the patients for the SSI and non-SSI populations can be seen in Table 3, along with the significance of the listed features in

distinguishing between the two populations. The prevalence rate of SSI in our data was 3.3% and there was a total of 38783 surgical encounters.

4.6.1 Model with Features from Statistical Feature Selection

This univariate filter followed by a multivariate filter was implemented as a means to reduce the feature set and aid in model selection. As described in Chapter 2, the univariate KS test provides evidence for the existence of correlation between the features and the binary target. First, the KS test filtered out features not correlated with the SSI target variable and yielded 206 features. Now, the cross-correlation among these features, known as Pearson's correlation coefficient, was calculated and for every pair of features with a coefficient value more than 0.2, one of the features from the pair was dropped since it was a redundant feature. This second filter yielded 27 features (See Appendix).

As mentioned previously, four ML models were trained using these features and their performances were evaluated. Table 2 shows the nested cv performance of the four ML models at a sensitivity value of 0.6. There were no significant differences in the performances, but RF, with a balance of slightly better performance and better computational complexity than XgBoost and AdaBoost, was selected for further experiments.

Table 2. Nested cross-validation performance of four models

Model	AUC	Sensitivity	PPV	NPV	Specificity
Random Forest	0.822 ± 0.02	0.6	0.14 ± 0.04	0.98 ± 0.001	0.86 ± 0.03
Logistic regression	0.821 ± 0.02	0.6	0.13 ± 0.03	0.98 ± 0.001	0.85 ± 0.03
XgBoost	0.825 ± 0.02	0.6	0.13 ± 0.03	0.98 ± 0.001	0.85 ± 0.03
AdaBoost	0.800 ± 0.02	0.6	0.12 ± 0.02	0.98 ± 0.001	0.84 ± 0.03

Table 3. Baseline characteristics and demographics of patients included in the study

Characteristics	SSI (n = 1316)	Non-SSI (n = 37207)	p-value
Patient			
Age (Median [IQR])	56 (23)	57 (25)	0.34
BMI (Median [IQR])	27.7 (8.3)	28.1 (8.3)	0.009
Female (n [%])	597 (45.4)	20591 (55.3)	<0.001
Male (n [%])	719 (54.6)	16616 (44.6)	<0.001
ASA physical status score			
≤ 2 (n [%])	212 (16.1)	13948 (37.5)	<0.001
> 2 (n [%])	1104 (83.9)	23259 (62.5)	<0.001
Surgical Procedure			
Exploratory Abdominal (n [%])	251 (19.0)	2392 (6.4)	<0.001
Small Bowel (n [%])	155 (11.8)	1060 (2.8)	<0.001
Colon (n [%])	125 (9.5)	876 (2.3)	<0.001
Craniotomy (n [%])	39 (3.0)	1507 (4.0)	0.73
Surgery risk			
low risk (n [%])	343 (26.0)	16587 (44.6)	<0.001
medium risk (n [%])	276 (21.0)	9786 (26.3)	0.05
high risk (n [%])	697 (53.0)	10834 (29.1)	<0.001
Labs - pre surgery			
Range of Platelet count (Median [IQR])	301.5 (281.5)	93.0 (147.0)	<0.001
Mean Albumin level (Median [IQR])	2.93 (1.34)	3.7 (0.44)	<0.001
Minimum haemoglobin level (Median [IQR])	7.3 (2.6)	10.0 (3.6)	<0.001
Mean Platelet count (Median [IQR])	269.0 (158.36)	225.66 (84.64)	<0.001
Minimum potassium level (Median [IQR])	3.2 (0.4)	3.5 (0.5)	<0.001
Minimum Albumin level (Median [IQR])	2.2 (1.6)	3.3 (0.8)	<0.001
Medications			
Diagnostic (n [%])	1157 (87.9)	20088 (54.0)	<0.001
Antihistamines (n [%])	1176 (89.4)	26639 (71.6)	<0.001
Antifungals (n [%])	713 (54.2)	7769 (20.9)	<0.001
Antidotes (n [%])	1046 (79.5)	17259 (46.4)	<0.001
Sedative hypnotics (n [%])	638 (48.5)	8476 (22.8)	<0.001
Anti- infectives (n [%])	124 (9.4)	1021 (2.7)	<0.001
Vitals			
Temperature difference before and after surgery (Median [IQR])	0.0 (0.9)	0.0 (0.0)	<0.001
Pulse range (Median [IQR])	64.0 (38.0)	38.0 (26.0)	<0.001
Max pulse (Median [IQR])	124.0 (34.0)	101.0 (27.0)	<0.001

4.6.2 Model with Features from MRMR and Forward Selection

Utilizing the feature selection technique presented in Chapter 2, 14 features that contributed to the highest-performing models were identified. This set contained features from different sources such as the laboratory values, the vitals, the medications, and surgery-related variables.

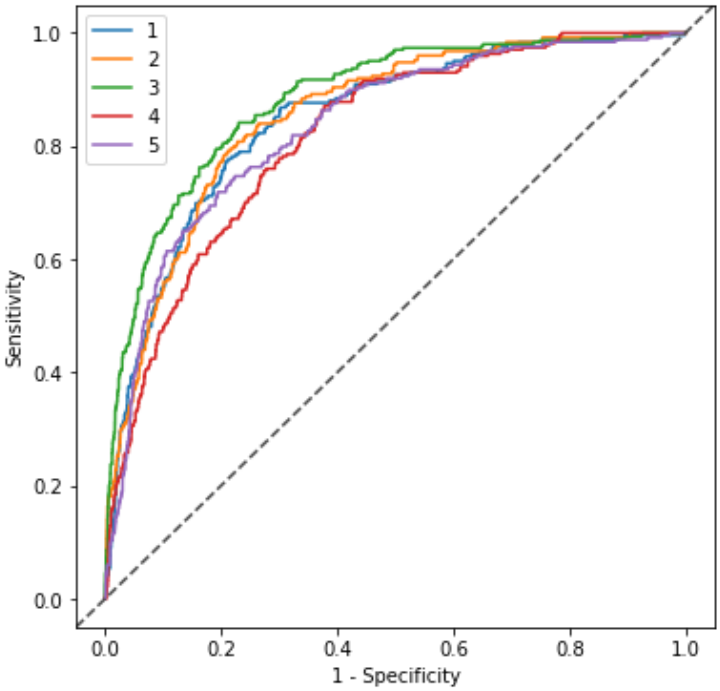

Figure 6. ROC curve for the 5-folds

Through our experiment, we discovered that none of the top features were from the post-surgical phase, which motivated us to build models using only the data from the pre-surgical phase to predict the risk of SSI. Feature selection was repeated on just the pre-surgery data, and the results mirrored the findings from our previous models; that most of the features were the same as the previous experiment. The performance of the RF model trained on these selected features can be seen in Table 4. The PPV, NPV, and the specificity were evaluated at four arbitrary values (0.6, .0.7, 0.8, 0.9) of sensitivity indicating four operating thresholds. Figure 6 and Figure 7 display

the ROC curves and precision-recall curves corresponding to the five folds. The following paragraphs discuss the features obtained from the pre-surgery data.

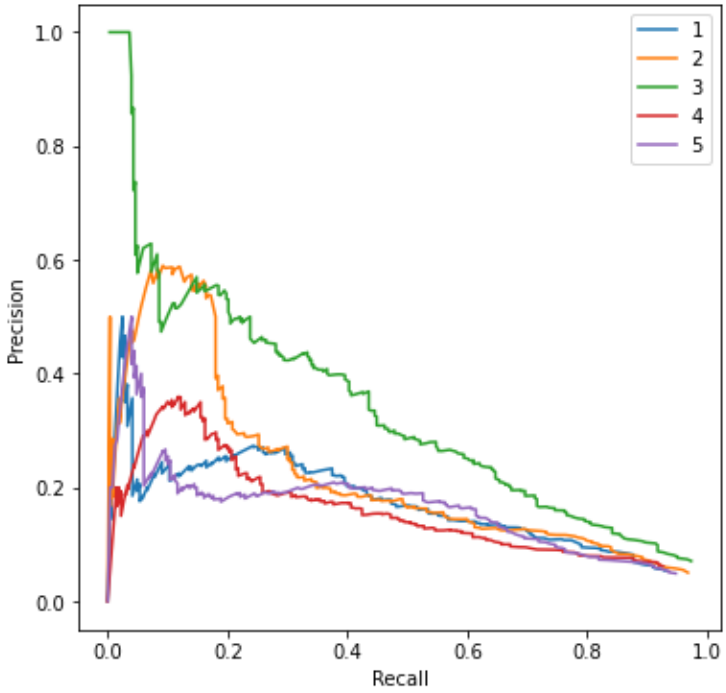

Figure 7. Precision-Recall curve for 5-folds

The laboratory variables included the range of platelet count prior to surgery which is defined as the difference between the maximum and minimum platelet count of a patient for a surgical procedure in the chosen period. This derived feature from the raw laboratory platelet values was a better predictor than the minimum, maximum, and mean values of platelet count, indicating that the range of the values provided more information than the absolute lab values by capturing the temporality. Other lab variables included, in the order of importance, the mean albumin level prior to surgery, the minimum hemoglobin level prior to surgery, minimum potassium level prior to surgery. The maximum pulse level prior to surgery was the only vitals variable present in the top 14. The medications variables were categorical variables indicating if the patient was administered the medication. The classes of medication variables selected were,

diagnostic, antifungals, sedative, anti-infectives, antidotes, and antihistamines. The surgery procedure variables were also categorical variables indicating the site of the surgical procedure and included sites such as “exploratory abdominal”, and “small bowel”. The top features also contained an expert-driven variable that quantified the level of risk for a patient into ‘low’, ‘mid’ and ‘high’, based on the type/site of the surgical procedure that was performed.

Table 4. Performance evaluation of RF and CNN trained using data from 5 days prior to the surgery.

Model	AUC	Sensitivity	PPV	NPV	Specificity
RF	0.85 ± 0.02	0.6	0.16 ± 0.05	0.98 ± 0.001	0.88 ± 0.03
		0.7	0.13 ± 0.03	0.99 ± 0.001	0.82 ± 0.04
		0.8	0.10 ± 0.02	0.99 ± 0.001	0.73 ± 0.05
		0.9	0.07 ± 0.02	0.99 ± 0.0003	0.60 ± 0.04
CNN	0.86 ± 0.01	0.6	0.16 ± 0.01	0.98 ± 0.001	0.89 ± 0.01
		0.7	0.14 ± 0.01	0.99 ± 0.001	0.85 ± 0.02
		0.8	0.11 ± 0.01	0.99 ± 0.001	0.77 ± 0.03
		0.9	0.08 ± 0.01	0.99 ± 0.001	0.63 ± 0.04

A similar trend about top-performing features was observed when we used a data set that included data from 5 days prior to surgery to 2 days after surgery, with the difference in temperature before and after surgery as an additional vitals’ variable.

The performance of CNN trained on the pre-surgery data can be seen in Table 4. Since neural networks can handle high dimension data, the input to the network was the entire feature set. CNNs are inherently capable of extracting important features as they are trained. The CNN model’s performance seems comparable to that of the RF model without an explicit feature selection process.

CHAPTER 5: BIAS INVESTIGATION

5.1 Setup

In this chapter, we describe the methods employed to identify the bias in the SSI data, and furthermore demonstrate the mitigation of this bias to avoid unfair disparity in decision making. The important features obtained from the previous chapters were used in training the RF model.

5.2 Bias Identification and Mitigation

In our experiment, the subgroups were defined based on sensitive attributes such as patient race and sex. The four subgroups defined were white-male, white-female, non-white-male, and non-white-female. The dataset was split into 80% training data and 20% hold-out test data, while maintaining the proportion of subgroup population as well as the proportion of positive and negative cases same across the two splits. Table 5 shows the proportion of the patient population for each subgroup in the training and hold-out datasets. The Random Forest algorithm was used to train the model on the entire training dataset and then evaluated on each subgroup hold-out dataset

A post-processing algorithm, as the name suggests, manipulates the posterior probabilities of the classifier, and changes the predicted label outputs to satisfy the fairness constraints. We demonstrate a simple post-processing algorithm that deals with choosing the right classification threshold for each subgroup such that the calculated metric satisfies the statistical constraint [27]. In other words, we aim to choose a fixed classification threshold for each subgroup such that the threshold-based metric score is equal for all the subgroups. By choosing different operating thresholds for each subgroup, we can deal with the offset in performance. In our study, we chose

Positive Predictive Value (PPV) as our metric since it was a sensible metric among the clinicians. Hence PPV parity was our definition of fairness which is defined as the equality in PPV across the subgroups. PPV is a metric derived from the confusion matrix, which could be altered by varying the threshold. The maximum attainable PPV of the least performing subgroup was chosen and the threshold for each subgroup was selected such that the PPV of that subgroup is equal to that of the chosen PPV. This can also be extended to other metrics of choice, such as the True Positive Rate parity and False Positive Rate parity.

Table 5. Subgroup populations in the training dataset and holdout dataset

Subgroups	**Training data (n = 30863)**			**Holdout data (n = 7660)**		
	SSI	**Non-SSI**	**Total**	**SSI**	**Non-SSI**	**Total**
White Male	373	8635	9008 (29.2%)	88	2153	2241 (29.3%)
White Female	318	10786	11104 (36%)	99	2692	2791 (36.4%)
Non-White Male	225	4628	4853 (15.7%)	33	1200	1233 (16%)
Non-White Female	150	5748	5898 (19.1%)	30	1365	1395 (18.2%)

5.3 Results

On evaluating the trained model on the subgroups of the holdout datasets at a fixed operating threshold of 0.1, a disparity in the model performance to predict SSI was observed. The disparate impact was calculated as the difference between the overall performance and each individual subgroup's performance (see Table 6). The disparate impact is an indicator of the bias in the model's performance across subgroups. The highest performing groups is the "white and female"

subgroup and the least performing is the “non-white and male”. These results also correlate with the subgroup population, as shown in Table 5, in that the largest group is the best performing one.

Table 6. PPV (at a threshold of 0.1) across the subgroups and the disparate impact with the overall PPV as 0.17

	White/male	**White/female**	**Non-white/male**	**Non-white/female**
Positive Predictive Value (PPV)	0.20	0.23	0.10	0.11
Disparate impact	0.03	0.06	0.07	0.06

Using the simple post-processing technique explained in the previous section, we were able to identify different operating thresholds for each subgroup such that the PPV is equal across the subgroups (see Table 7). Operating the classifier at these individual subgroups thresholds satisfies our PPV fairness constraints. Figure 8 shows the PPV across different thresholds for the defined subgroups. A horizontal line at PPV = 0.19 would intersect all the curves at different points of operation. This similar technique can be used to realise other fairness constraints based on other statistical metrics of choice.

Table 7. Chosen thresholds for each subgroup such that the statistical parity constraint is satisfied.

	White/male	**White/female**	**Non-white/male**	**Non-white/female**
Threshold	0.09	0.08	0.26	0.16
Positive Predictive Value	0.19	0.19	0.19	0.19

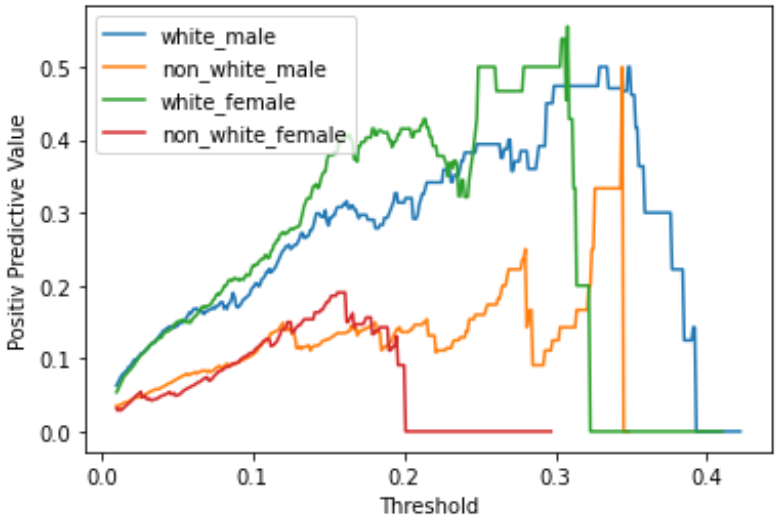

Figure 8. Positive Predictive Values (PPV) across different thresholds for the four subgroups

CHAPTER 6: DISCUSSION AND CONCLUSION

6.1 Discussion

Prediction, risk assessment, and early detection are important tools to reduce SSI, one of the most common, costly, yet potentially preventable adverse surgical outcomes. CDC initiated surveillance methods to track SSI and guidelines for preventative measures have been credited with improvements in SSI rates, though much more efforts are needed to further curtail the SSI burden. This national priority has spurred many research efforts, including accurate and advanced SSI prediction models and risk stratification, which are critical to alert the clinicians to take measures at risk factor modification, before, during, and after surgical procedures, including risk-based reconsiderations to surgical approaches.

The approaches and features (risk factors) considered at developing SSI prediction and risk stratification models are varied. Some are generalizable across all surgical procedures, some are organ system-specific (e.g., cardiothoracic, neurosurgery, orthopedic) and others are procedure-specific (colorectal resection, ventral hernia repair, lower extremity revascularization, hip replacement, spinal surgery). Some of the commonly studied risk factors include patient characteristics (e.g., age, gender, BMI, and smoking status), pre-operative biomarkers, including inflammatory markers (e.g., hemoglobin, WBC and platelet counts, CRP, albumin), co-morbidities, ASA classification of patient physical status, incision wound classification, complexity and/or duration of surgery, ambulatory or inpatient, and emergency or planned.

In this thesis, all the available structured data, routinely collected during patient care and monitoring, from multiple health information systems, such as institutional data warehouses and registries (NHSN), were extracted and processed. Data covered pre-, intra-, and postoperative

periods, up to 3 months after surgery. Data included all the NHSN registry variables, laboratory data, medications administered, and vitals collected during patient care. The data was a comprehensive set and covered the full continuum of SSI events from the planning stage and provided rich data encompassing most of the risk factors implicated in the literature, more than any other study had used for analysis and modeling. Six ML models, including Neural Networks, were tested for SSI classification, and based on the performance, the RF and the CNN algorithms were chosen for further experimentation.

Our data contained patients who underwent 38 different surgical procedures. Many studies included patients from a specific population with a certain risk level. They also developed algorithms that model surgery-specific populations rather than including a wide range of surgical procedures [33, 34, 35, 36].

There were very few studies that incorporated EHR data such as labs, vitals, and medications as opposed to a narrow focus on surgery and patient variables. But like many other approaches, the modeling approaches involved prediction in the postoperative phase, but the temporal information of the pre-operative data was not part of the study. There are not many studies that have used pre-operative biomarkers, and this study combined the routinely conducted pre-operative laboratory test results with patient characteristics, co-morbidities, and vitals to look at SSI prediction and risk stratification. Our approach utilizes the temporal behavior of the pre-operative laboratory, vitals, and other patient characteristics.

Based on the features selected from models including pre-operative and postoperative data, it was found that the most significant predictors were pre-operative data. This motivated us to conduct further analysis to include only the pre-operative data in features selection and predictive modeling. This yielded results with little to no change compared to the models built from including

postoperative data in the feature selection stage. This indicates that pre-operative laboratory, vitals, and medication data combined with certain surgery-related features can enable the early detection of SSI.

The effectiveness of Artificial Neural Networks (ANNs) in biomedical informatics has been witnessed in multiple studies in the past. Though ANNs are primarily used for unstructured data, efforts have been made to avail the ability of ANNs to model complex relationships in structured Electronic Medical Records (EMR) data [18, 19, 37]. In our study, we used a one-dimensional shallow Convolutional Neural Network (CNN) to model the relationship between the input values and the target values. About 290 scaled input features containing variables from different sources such as patient-related, surgery-related, laboratory, vitals, and medications were used. The performance of the CNN was observed to be similar to that of the Random Forest algorithm, if not better.

Further, on analyzing the fairness of these predictive models across different subgroups, we observed a disparity. There are multiple reasons for the observed disparity in performance. We know the algorithm is only as good as the data, and any imperfection in the data is inherited by the model. Data can also reflect the human and societal biases that cause them to be skewed or tainted. In most cases, it is hard to identify the reason for bias. The difference in sample size among the subgroups (See Table 5), as observed in our data, could also cause the algorithm to imperfectly model the minority groups. Due to the varying sample size, the model performance for the minority subgroups is sub-optimal with varying performance across the subgroups which can, in turn, lead to decisions that result in drastic surgical outcomes for the minority subgroups.

The post-processing method employed to mitigate the observed bias comes with certain limitations. First, it requires test time access to the sensitive attributes. Second, due to the trade-

off between accuracy and fairness, implementing a fair classifier means that the well-performing subgroups should settle for lower accuracies. The maximum accuracy achievable is based on the maximum value for the least performing group. Further, we would also like to indicate that there are developed toolkits from IBM [24], Google [38], and Facebook that implement different algorithms assisting in detecting and mitigating bias.

This experiment was performed to demonstrate the existence of biases in prediction and risk scoring applications that focus on critical study environments such as in clinical settings. We believe that it is necessary to acknowledge the existence of bias in ML models, that may impact the provision of appropriate preventative medical intervention to avoid SSI in the minority or the historically underrepresented groups and further to mitigate the propagation of the above biases into any future applications.

Our study also comes with certain limitations. First, the data was from a single institution and was internally validated. Second, some of the challenges of retrospective study are applicable to our study [39]. Further validation of our models with data from other institutions and prospective studies is necessary to confirm our findings. Third, better understanding and handling of missing data and imputation methods are required to reflect the real-world variance in data.

6.2 Conclusion

In this thesis, we report a study of SSI risk detection using various predictive modeling algorithms on a patient cohort containing multiple surgical procedures and many features. The important risk factors for the early detection of SSI were identified. And the potential application of Convolutional Neural Networks in structured EMR data for the prediction of SSI risk was demonstrated. We also identified the existence of the issue of bias in ML modeling used to predict SSI and demonstrated a method to mitigate the bias.

REFERENCES

[1] Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

[2] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[3] Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253), 68–78. https://doi.org/10.2307/2280095

[4] (2008) Pearson's Correlation Coefficient. In: Kirch W. (eds) Encyclopedia of Public Health. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-5614-7_2569

[5] Zhao, Z., Anand, R., & Wang, M. (2019). Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 442-452.

[6] Anderson DJ, Podgorny K, Berríos-Torres SI, Bratzler DW, Dellinger EP, Greene L, Nyquist AC, Saiman L, Yokoe DS, Maragakis LL, Kaye KS. Strategies to prevent surgical site infections in acute care hospitals: 2014 update. Infect Control Hosp Epidemiol. 2014 Jun;35(6):605-27. doi: 10.1086/676022. PMID: 24799638; PMCID: PMC4267723.

[7] Steiner CA, Karaca Z, Moore BJ, Imshaug MC, Pickens G. Surgeries in Hospital-Based Ambulatory Surgery and Hospital Inpatient Settings, 2014: Statistical Brief #223. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs. Rockville (MD), 2006.

[8] Zimlichman E, Henderson D, Tamir O, et al. Health care-associated infections: a meta-analysis of costs and financial impact on the US health care system. JAMA Intern Med 2013;173(22):2039-46 doi: 10.1001/jamainternmed.2013.9763[published Online First: Epub Date].

[9] Price C, Savitz L. Improving the measurement of surgical site infection risk stratification/outcome detection. Final report (Prepared by Denver Health and its partners under Contract No. 290 2006 00 20). AHRQ Publication 2012(12):0046

[10] National Healthcare Safety Network (NHSN). Secondary National Healthcare Safety Network (NHSN). https://www.cdc.gov/nhsn/index.html.

[11] National Surgical Quality Improvement Program (NSQIP). Secondary National Surgical Quality Improvement Program (NSQIP). https://www.facs.org/quality-programs/acs-nsqip

[12] Korol E, Johnston K, Waser N, Sifakis F, Jafri HS, Lo M, Kyaw MH. A systematic review of risk factors associated with surgical site infections among surgical patients. PLoS One. 2013 Dec 18;8(12):e83743. doi: 10.1371/journal.pone.0083743. PMID: 24367612; PMCID: PMC3867498.

[13] Ke C, Jin Y, Evans H, Lober B, Qian X, Liu J, Huang S. Prognostics of surgical site infections using dynamic health data. J Biomed Inform. 2017 Jan;65:22-33. doi: 10.1016/j.jbi.2016.10.021. Epub 2016 Nov 5. PMID: 27825798.

[14] Mueck KM, Kao LS. Patients at High-Risk for Surgical Site Infection. Surg Infect (Larchmt). 2017 May/Jun;18(4):440-446. doi: 10.1089/sur.2017.058. Epub 2017 Apr 12. PMID: 28402740.

[15] Hu Z, Simon GJ, Arsoniadis EG, Wang Y, Kwaan MR, Melton GB. Automated Detection of Postoperative Surgical Site Infections Using Supervised Methods with Electronic Health Record Data. Stud Health Technol Inform. 2015;216:706-10. PMID: 26262143; PMCID: PMC5648590.

[16] Fletcher RR, Olubeko O, Sonthalia H, Kateera F, Nkurunziza T, Ashby JL, Riviello R, Hedt-Gauthier B. Application of Machine Learning to Prediction of Surgical Site Infection. Annu Int Conf IEEE Eng Med Biol Soc. 2019 Jul;2019:2234-2237. doi: 10.1109/EMBC.2019.8857942. PMID: 31946345.

[17] Jiang Z, Ardywibowo R, Samereh A, Evans HL, Lober WB, Chang X, Qian X, Wang Z, Huang S. A Roadmap for Automatic Surgical Site Infection Detection and Evaluation Using User-Generated Incision Images. Surg Infect (Larchmt). 2019 Oct;20(7):555-565. doi: 10.1089/sur.2019.154. Epub 2019 Aug 19. PMID: 31424335; PMCID: PMC6823883.

[18] Walczak, Steven & Davila, Marbelly & Velanovich, Vic. (2019). Prophylactic antibiotic bundle compliance and surgical site infections: An artificial neural network analysis. Patient Safety in Surgery. 13. 10.1186/s13037-019-0222-4.

[19] Adavi M, Salehi M, Roudbari M. Artificial neural networks versus bivariate logistic regression in prediction diagnosis of patients with hypertension and diabetes. Med J Islam Repub Iran. 2016 Jan 3;30:312. PMID: 27390682; PMCID: PMC4898876.

[20] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[21] Lohia, Pranay Kr. et al. “Bias Mitigation Post-processing for Individual and Group Fairness.” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2019): 2847-2851.

[22] Gajane, P. (2017). On formalizing fairness in prediction with machine learning. *ArXiv, abs/1710.03184*.

[23] S. Verma and J. Rubin, "Fairness Definitions Explained," 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), 2018, pp. 1-7, doi: 10.23919/FAIRWARE.2018.8452913.

[24] Bellamy, R. et al. “AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.” *ArXiv* abs/1810.01943 (2018): n. Pag.

[25] Barocas, Solon, and Andrew D. Selbst. "Big Data's Disparate Impact." California Law Review 104, no. 3 (2016): 671-732.

[26] d'Alessandro, Brian & O'Neil, Cathy & LaGatta, Tom. (2017). Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. Big Data. 5. 120-134. 10.1089/big.2016.0048.

[27] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[28] Wickham, Hadley & Averick, Mara & Bryan, Jennifer & Chang, Winston & McGowan, Lucy & François, Romain & Grolemund, Garrett & Hayes, Alex & Henry, Lionel & Hester, Jim & Kuhn, Max & Pedersen, Thomas & Miller, Evan & Bache, Stephan & Müller, Kirill & Ooms, Jeroen & Robinson, David & Seidel, Dana & Spinu, Vitalie & Yutani, Hiroaki. (2019). Welcome to the Tidyverse. Journal of Open Source Software. 4. 1686. 10.21105/joss.01686.

[29] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

[30] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. ArXiv, abs/1912.01703.

[31] van Walraven C, Musselman R. The Surgical Site Infection Risk Score (SSIRS): A Model to Predict the Risk of Surgical Site Infections. PLoS One. 2013 Jun 27;8(6):e67167. doi: 10.1371/journal.pone.0067167. PMID: 23826224; PMCID: PMC3694979.

[32] Akiba, Takuya et al. “Optuna: A Next-generation Hyperparameter Optimization Framework.” *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019): n. Pag.

[33] Li X, Nylander W, Smith T, Han S, Gunnar W. Risk Factors and Predictive Model Development of Thirty-Day Post-Operative Surgical Site Infection in the Veterans Administration Surgical Population. Surg Infect (Larchmt). 2018 Apr;19(3):278-285. doi: 10.1089/sur.2017.283. Epub 2018 Feb 1. PMID: 29389252.

[34] Berger RL, Li LT, Hicks SC, Davila JA, Kao LS, Liang MK. Development and validation of a risk-stratification score for surgical site occurrence and surgical site infection after open ventral hernia repair. J Am Coll Surg. 2013 Dec;217(6):974-82. doi: 10.1016/j.jamcollsurg.2013.08.003. Epub 2013 Sep 17. PMID: 24051068.

[35] Gervaz P, Bandiera-Clerc C, Buchs NC, Eisenring MC, Troillet N, Perneger T, Harbarth S. Scoring system to predict the risk of surgical-site infection after colorectal resection. Br J Surg. 2012 Apr;99(4):589-95. doi: 10.1002/bjs.8656. Epub 2012 Jan 9. PMID: 22231649.

[36] Kuy S, Dua A, Desai S, Dua A, Patel B, Tondravi N, Seabrook GR, Brown KR, Lewis BD, Lee CJ, Kuy S, Subbarayan R, Rossi PJ. Surgical site infections after lower extremity revascularization procedures involving groin incisions. Ann Vasc Surg. 2014 Jan;28(1):53-8. doi: 10.1016/j.avsg.2013.08.002. Epub 2013 Nov 1. PMID: 24189008.

[37] Kuo PJ, Wu SC, Chien PC, Chang SS, Rau CS, Tai HL, Peng SH, Lin YC, Chen YC, Hsieh HY, Hsieh CH. Artificial neural network approach to predict surgical site infection after free-flap reconstruction in patients receiving surgery for head and neck cancer. Oncotarget. 2018 Feb 9;9(17):13768-13782. doi: 10.18632/oncotarget.24468. PMID: 29568393; PMCID: PMC5862614.

[38] D'Amour, A. et al. "Fairness is not static: deeper understanding of long term fairness via simulation studies." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020): n. Pag.

[39] Hess DR. Retrospective studies and chart reviews. Respir Care. 2004 Oct;49(10):1171-4. PMID: 15447798.

[40] Bscan, (March 2013). Illustration of the Kolmogorov–Smirnov statistic. Retrieved from https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

[41] Steven Nickolas, Feb 2019. Correlation coefficient. Retrieved from https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

APPENDIX A

Table A.1. Important features derived from data spanning 30 days prior to 7 days after surgical procedure

Range of platelet count prior to surgery
Mean Albumin level prior to surgery
Diagnostic medication prior to surgery
Risk of surgery
Minimum haemoglobin level prior to surgery
Maximum pulse measurement prior to surgery
Antifungal medication prior to surgery
Exploratory abdominal surgical procedure
Small bowl surgical procedure
Sedative medication prior to surgery
Minimum albumin level prior to surgery
Colon surgical procedure
Anti-infective medication prior to surgery
Craniotomy surgical procedure

Table A.2. Important features derived from data 5 days prior to surgical procedure

Range of platelet count prior to surgery
Mean Albumin level prior to surgery
Minimum haemoglobin level prior to surgery
Risk of surgery
Diagnostic medication prior to surgery
Antidotes medication prior to surgery
Antifungals medication prior to surgery
Maximum pulse measurement prior to surgery

Antihistamines medication prior to surgery
Exploratory abdominal surgical procedure
Anti-infective medication prior to surgery
Minimum potassium level prior to surgery
Small bowl surgical procedure
Sedative medication prior to surgery

Table A.3. Important features derived from data spanning 5 days prior to 2 days after surgical procedure.

Range of platelet count prior to surgery
Mean Albumin level prior to surgery
Diagnostic medication prior to surgery
Minimum haemoglobin level prior to surgery
Risk of surgery
Antihistamines medication prior to surgery
Difference in temperature before and after surgery
Range of pulse measurement prior to surgery
Antifungals medication prior to surgery
Exploratory abdominal surgical procedure
Small bowl surgical procedure
Colon surgical procedure
Antidotes medication prior to surgery
Mean platelet count prior to surgery