Final Report,
IDC MLDS MSc

Date: Oct/2021

Title: **Predicting Anastomotic Leakage Risks In Gastro Surgeries**

Student: Maya Kerem
Supervisor: Prof Zohar Yakhini
Collaborators: Nomi Hadar, Dr. Dan Assaf, Dr. Iris Shahar

Abstract

Anastomotic Leakage is a major and potentially deadly complication that may arise after abdominal surgeries [1]. It is the most frequent and serious post-operative complication of colorectal surgery and results in mortality rates up to 40%. The complication occurs when luminal content leaks from the surgical join.

Using data from Sheba Hospital, consisting of 524 patients, where 8.21% of the patients developed Anastomotic Leakage, we use statistical analysis tools and Naïve Bayes [2][3] models in order to predict the development of AL in patients. We demonstrate a machine learning model that allows us to capture 90% of all AL patients (recall = 0.9) with a total positive rate of 40%. In other words, we attain 0.9 recall at a total alert rate (TAR) of 0.4.

Our work is an initial step towards potentially assisting practitioners in managing treatments for abdominal surgery patients and avoiding post-operative complications.

Introduction

AL is the most frequent and serious post-operative complication of colorectal surgery.
AL occurs when the intestinal wall is damaged and causes leakage of colonic content into the abdominal or pelvic cavity. AL is a dangerous complication with mortality rates up to 40%.

There are several ways by which colorectal/abdominal surgery patient management can benefit from accurate predictive models. First, patients with high AL risk can be further monitored and assigned for longer post-operative hospitalization as a preventative measure. In addition, variations in the surgical procedure can be considered. For example, temporary diverting stoma might be a protective procedure for the prevention of anastomotic leakage (AL) after anterior resection.

Diverting stoma, placed in the proximal colon or ileum during the initial operation, prevents anastomotic leakage and clearly reduces the incidence of peritonitis and thus re-operations, and mortality [4]. Thus, for patients with predicted high anastomotic leakage risk, the operating surgeon might decide to construct a diverting stoma.

The purpose of this work is to provide initial predictive models as a proof of concept in the context of assisting practitioners in managing treatments for abdominal surgery patients.

Our data consists of 524 patients from Shiba Medical Center Hospital including measured data points obtained from medical history, hospital measurements and tests. The considered medical history parameters included are gender, age, surgical approach and surgery agency while hospital measurements include, for example, Albumin levels, Calcium levels, weight, BMI and more.

The data was collected both before and after the actual surgery, these data will be referred to as Pre Surgery data and Post Surgery data respectively.

Data preparation steps included KNN imputation and outlier correction as well as multiple statistical evaluations, such as Wilcoxon Rank Sum [5] and Fisher's Exact Test [6]. Our machine learning approach was based on Naïve Bayes models. We considered three different feature types: Features from Pre Surgery, features from Post surgery and features combining before and after surgery information, referred to as Combined Surgery data. In addition, we consider three different feature selection procedures: Vanilla model (no feature selection), Disjointified features model [7], and Disjointified features model after FDR [8]. The developed Naïve Bayes models, given any fixed threshold, output a binary prediction: 0 for no AL and 1 for AL.

An interesting aspect of the evaluation methodology used herein is the calculation of stay-in-hospital rate (SIH, see Methods), also referred to Total Alert Rate (TAR), for every relevant Bayesian threshold. In particular, we examine the SIH at different desired levels of recall. We developed visualization tools that allow for plotting SIH against other evaluation performance characteristics such as ROC curves and PR curves. These tools will be available as part of the code deployed to GitHub.

We show that a predictive model, using only data from pre surgery, can capture 90% of all AL patients (recall = 0.9) with a total positive rate of 40% (TAR = SIH = 0.4). In other words, we attain 0.9 recall when alerting only 40% of the patients.

Our work, including the predictive models and the methodology, constitutes a potentially important step towards efficient treatment management in abdominal surgery patients. Future work addressing larger dataset will hopefully validate and refine our findings.

Results

The data used was in the form of measured events in addition to some basic patient parameters. The patient parameters were Gender, Age, Surgical approach (laparoscopic vs open) and surgery urgency. There were 53 measured event parameters such as: blood pressure, weight, Albumin levels, Calcium levels, Potassium Levels and more. Event parameters with multiple measurements were summarized as described in Methods.

Below we describe the analysis and modeling results. This is all done for a 33% test data with a random split. In addition, we report 5-fold cross validation results in the Appendix. Details of the methodology are further described in Methods.

Pre-Surgery Data

Our first group of Naïve Bayes models used pre-surgery data. To statistically evaluate the features, we used WRS. Figure 1 depicts the results. We observe 12 features at an FDR q=0.15 (Figure 2)

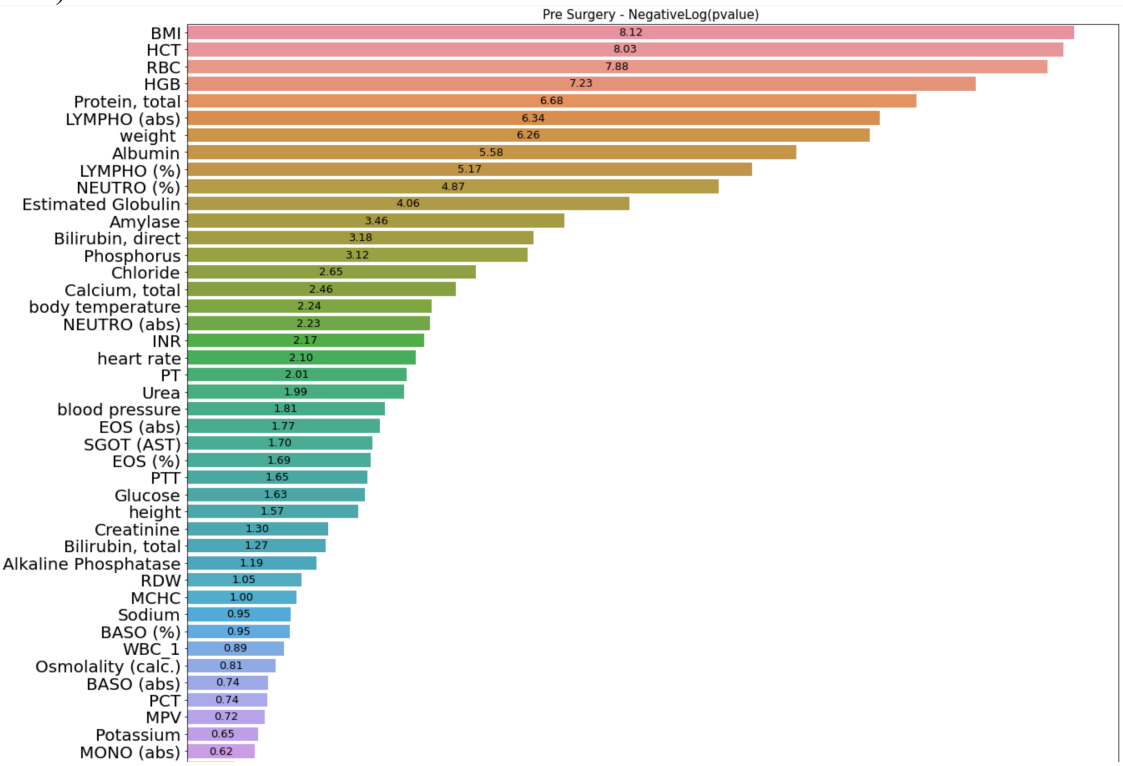


Figure 1 - negative log WRS p-value for pre-surgery features, calculated for the AL label

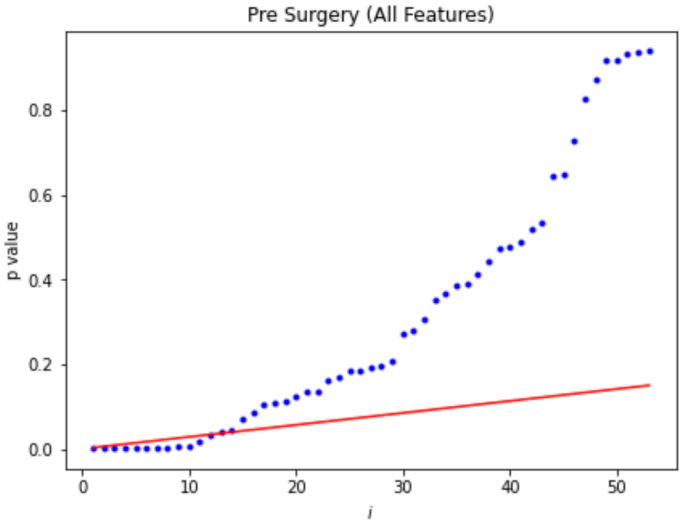


Figure 2 - false discovery rate (FDR) for all features prior to surgery with q = 0.15, depicted as the red line

The first Naïve Bayes model, labeled 'vanilla', considered all 55 features. The second Naïve Bayes model considered the set of features after a disjointification process ranking by WRS p-values and a Pearson threshold of 0.5, resulting in 38 features. The last Naïve Bayes model uses a set of features obtained by applying FDR with q=0.1 to the disjointified set of features (same parameters), resulting in 6 features. Note that this process was run using the training set only. We call these models VNB, DNB and FDNB, respectively.

Figure 3.A depicts ROC curves for the three feature selection approaches. The largest AUC is attained by the combined approach: Disjointified and FDR.

When comparing the three feature selection approaches at a recall of 0.9, we note that the Vanilla model attains an SIH rate of 0.4, the Disjointified model attains an SIH rate of 0.56 and the Disjointified with FDR model attaining an SIH rate of 0.55. Seeking a minimal SIH at a recall of 0.9, we would select the Vanilla model representing optimal potential performance. (Figure 3.B-D)

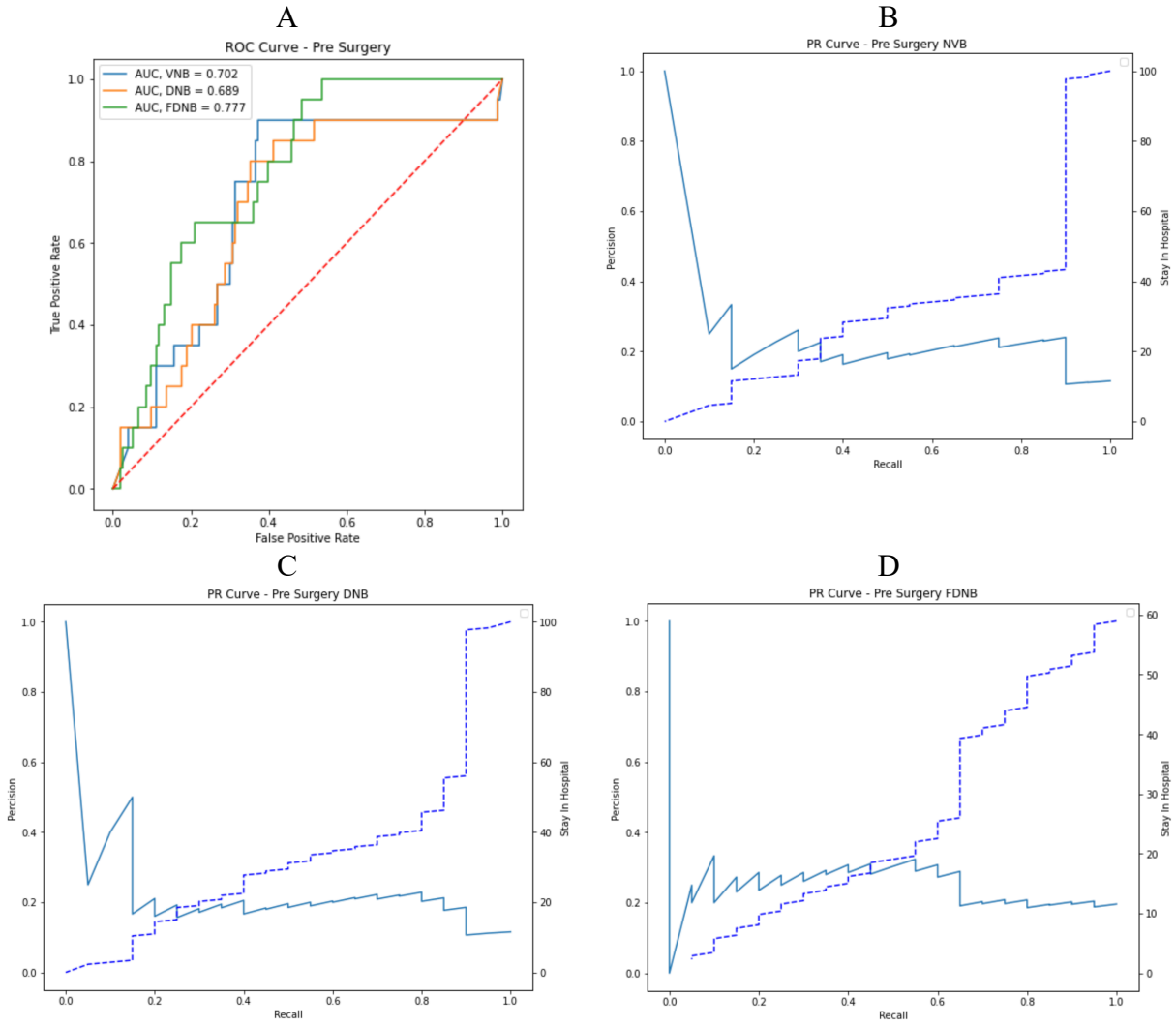


Figure 3 - Panel A: ROC curve for all three feature selection approaches, pre-surgery data. Panel B-D: PR curves with relevant SIH for the three feature selection approaches, pre-surgery data

Post-Surgery Data

The second group of Naïve Bayes models used post-surgery data. To statistically evaluate the features, we used WRS. Figure 4 depicts the results. We observe 30 features at an FDR q=0.01 (Figure 5)

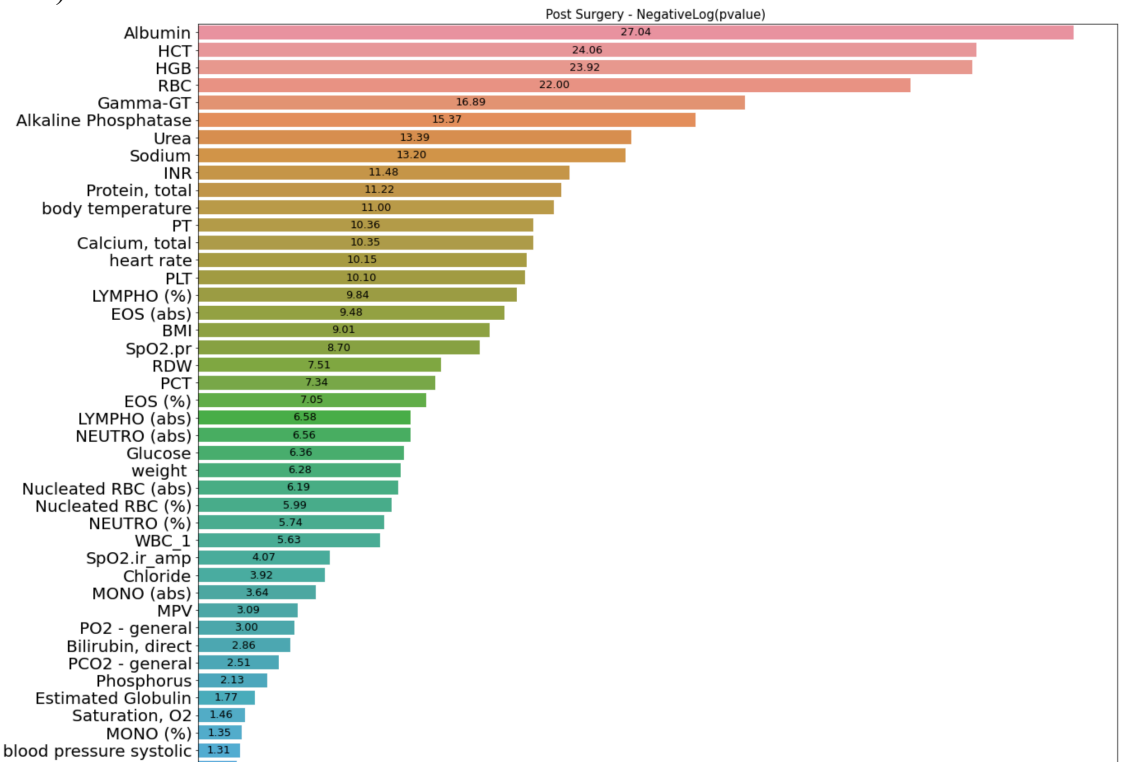


Figure 4 - negative log WRS p-value for post-surgery features, calculated for the AL label

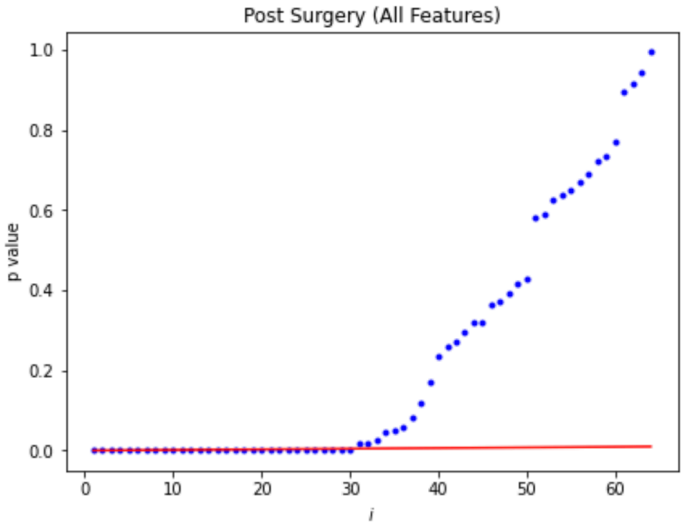


Figure 5 - false discovery rate (FDR) for all features post-surgery with q = 0.01, depicted as the red line

The first Naïve Bayes model, labeled 'vanilla', considered all 66 features. The second Naïve Bayes model considered the set of features after a disjointification process ranking by WRS p-values and a Pearson threshold of 0.5, resulting in 36 features. The last Naïve Bayes model uses

a set of features obtained by applying FDR with q=0.05 to the disjointified set of features (same parameters), resulting in 16 features. Note that this process was run using the training set only. We call these models VNB, DNB and FDNB, respectively.

Figure 6.A depicts ROC curves for the three feature selection approaches. The largest AUC is attained by the combined approach: Disjointified and FDR.

When comparing the three feature selection approaches at a recall of 0.9, we note that the Vanilla model attains an SIH rate of 0.47, the Disjointified model attains an SIH rate of 0.4 and the Disjointified with FDR model attains an SIH rate of 0.45. Seeking a minimal SIH at a recall of 0.9, we would select the Disjointified model representing optimal potential performance. (Figure 6.B-D)

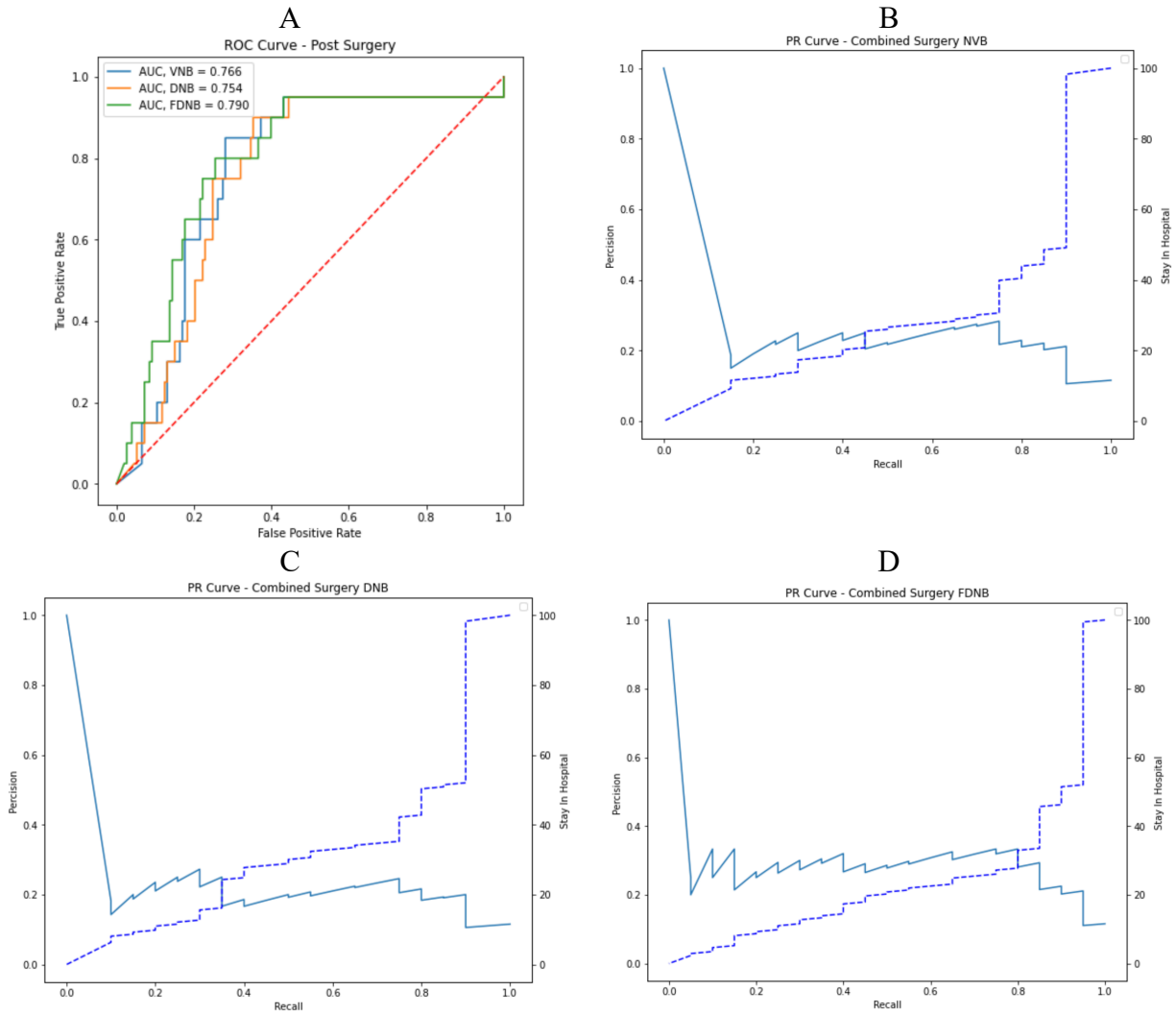


Figure 6 – Panel A: ROC curve for all three feature selection approaches, post-surgery data. Panel B-D: PR curve with relevant SIH for the three feature selection approaches, post-surgery data

Combined Data

The third group of Naïve Bayes models used pre-surgery and post-surgery data together. To statistically evaluate the features, we used WRS. Figure 7 depicts the results. We observe 36 features at an FDR q=0.01 (Figure 8)

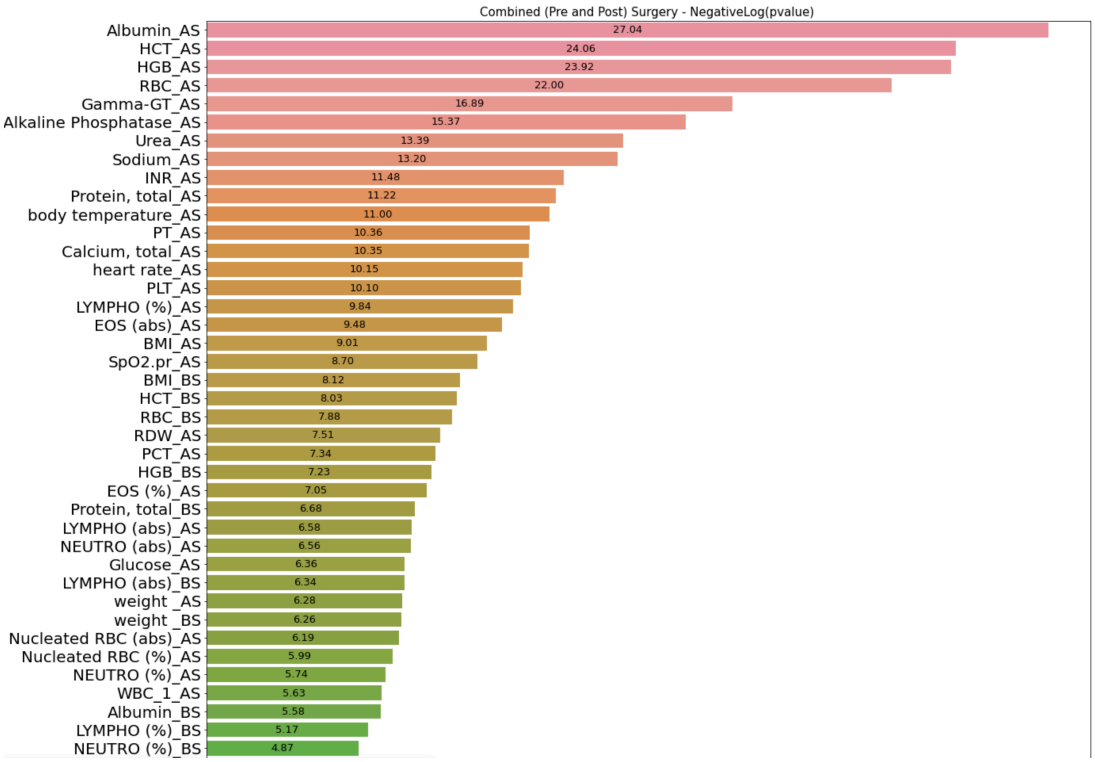


Figure 7 - negative log WRS p-value for combined-surgery features, calculated for the AL label

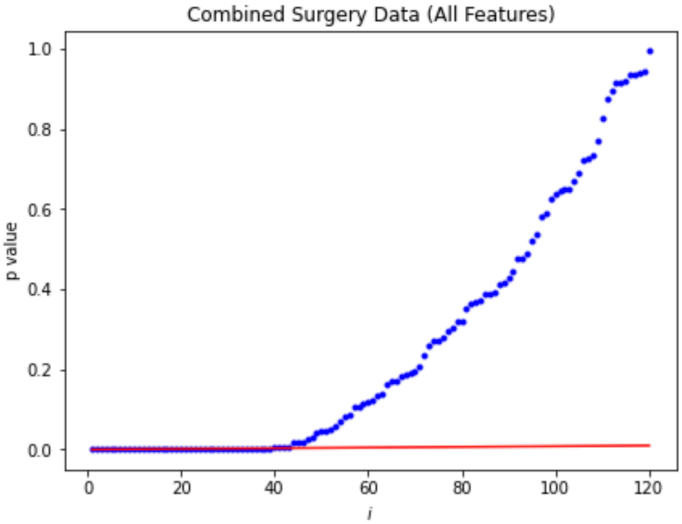


Figure 8 - false discovery rate (FDR) for all features combined-surgery with q = 0.01, depicted as the red line

The first Naïve Bayes model, labeled ‘vanilla’, considered all 123 features. The second Naïve Bayes model considered the set of features after a disjointification process ranking by WRS p-values and a Pearson threshold of 0.5, resulting in 70 features. The last Naïve Bayes model uses a set of features obtained by applying FDR with q=0.05 to the disjointified set of features (same parameters), resulting in 18 features. Note that this process was run using the training set only. We call these models VNB, DNB and FDNB, respectively.

Figure 9.A depicts ROC curves for the three feature selection approaches. The largest AUC is attained by the combined approach: Disjointified and FDR.

When comparing the three feature selection approaches at a recall of 0.9, we note that the Vanilla model attains an SIH rate of 0.5, the Disjointified model attains an SIH rate of 0.45 and the Disjointified with FDR model attains an SIH rate of 0.4. Seeking a minimal SIH at a recall of 0.9, we would select the Disjointified with FDR model representing optimal potential performance. (Figure 9.B-D)

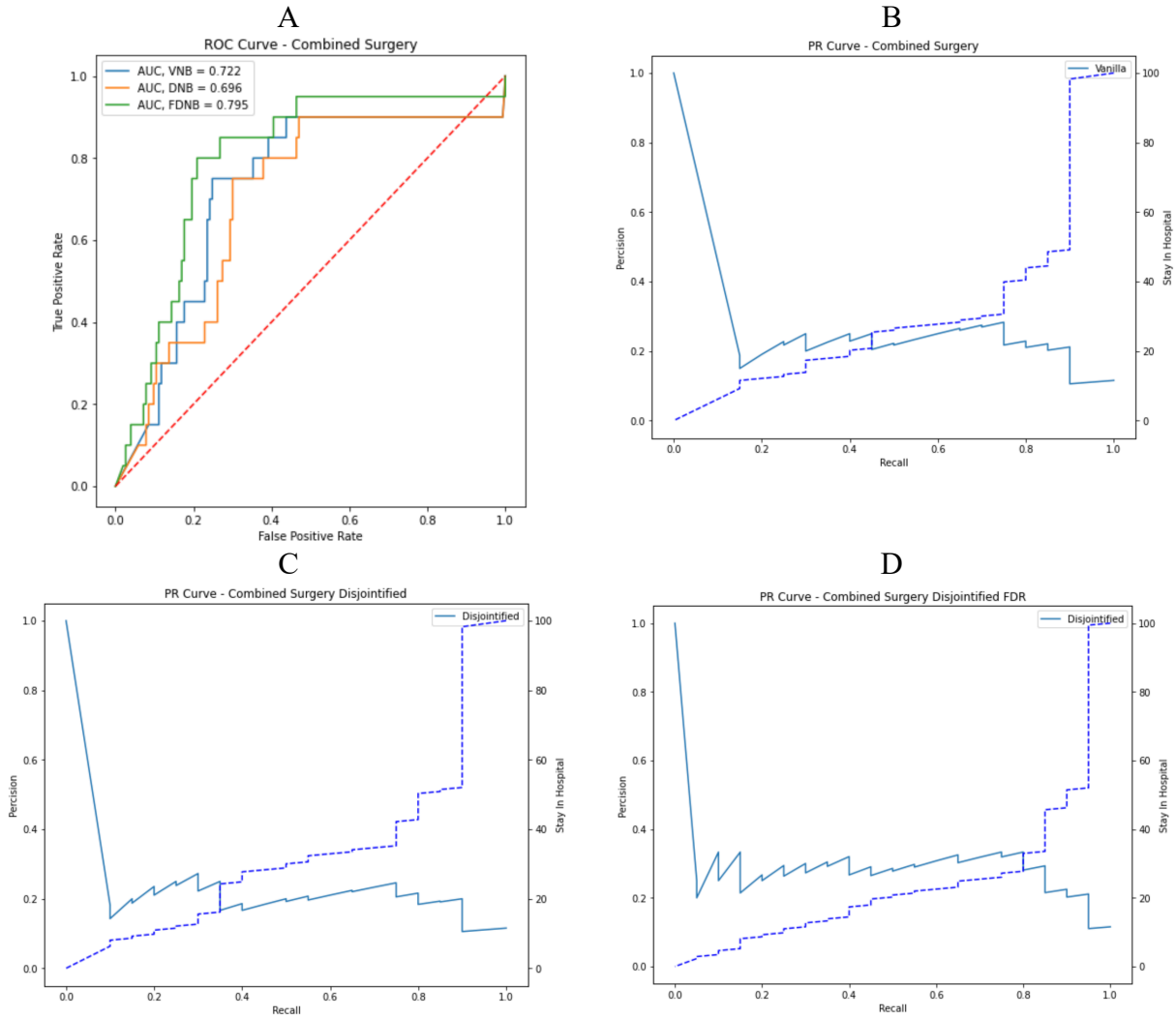


Figure 9 – Panel A: ROC curve for all three feature selection approaches, combined-surgery data. Panel B-D: PR curve with relevant SIH for the three feature selection approaches, combined-surgery data

Methods

Feature engineering and data preparation

We started with patient level data, where for each patient we had a file describing all measured events. Additional patient data was contained in a separate file. This data was combined from all patients to form one matrix of 525 patients and 481 features. Hospital measurements were acquired up to 4 times per patients, in such cases the mean value was considered. When splitting the data to train and test, we used 33% of the data as test data. We also report findings from 5-fold cross validation, see Appendix.

In order to handle missing data and to correct outliers we used kNN imputation with $k = 5$. Imputation was performed on the training set only. For test samples, any missing values were imputed by using the training mean for that feature.

FDR (features statistics)

False Discovery Rate [8][9] is formally defined as the proportion of errors committed by falsely rejecting the null hypothesis and is calculated as follows:

$$FDR(i) = \frac{i}{N \cdot p(i)}$$

Where $p(i)$ is the p value of feature i for the list of features sorted by p value and N is the total number of features. Fixing a desired FDR level, q, we therefore aim, in principle, to use all the features of which p-value is less than $\frac{i}{N} \cdot q$.

Disjointification for feature selection

Disjointification is a heuristic iterative procedure designed to find correlated features that have a similar association to the target class, and then not including them in the selected set of final features for the model.

The computation uses the p-value of all features using WRS (Fisher's Exact Test for categorical data). We first rank all features according to their p-value. Running on this ranked list, we compute the Pearson correlations for the ith feature in the list and all features in the currently selected set. The ith feature is considered selected if this Pearson correlation is less than the threshold of $|0.5|$ for all currently selected features. Thus, a lower-ranked feature does not make it into the selected list of features when the latter already has a feature with a similar impact.

We present two pseudocode versions for Disjointification process:

```
features_list = sorted features list by p-value
for i in range(number of features):
    list = [top i from features_list]
    Calculate the correlation between ith feature and all
        remaining features from the list
    If correlation > threshold for any comparison
    Remove the ith feature from features_list
return features_list
```

```
feature[] = rank features (ascending) according to p-value, yielding a
sorted list of length L
selected_list = ∅  //Initialization
for i in range (L)
    compute the correlation of feature[i] to all features in
    selected_list
    if all correlations < τ, then append (feature[i], selected_list)
return selected_list
```

PR Curve

The PR curve shows the relationship between the precision and recall for each model per specific threshold. We used the library `sklearn.metrics.precision_recall_curve`

SIH (Total Alert Rate)

SIH, or Total Alert Rate, is the percent of patients that need to stay in the hospital in order to achieve a desired level of recall. This, of course, assumes, for the purpose of naming, that the clinical management step that follows an alert comprises continued hospitalization. Using i that runs through patients, sorted by the posterior value of Bayes model prediction, we compute:

$$SIH(i) = \frac{FP(i) + TP(i)}{Total\ Patients}$$

We can now plot SIH overlayed on to the PR curve. A good model will have low SIH rates for relatively high recall rates. To adjust to the limitations of sklearn we used the following formula where i is the threshold iterator index:

$$SIH(i) = \frac{(Recall(i) \cdot TotalPos)\left(\frac{\left(1 - Precision(i)\right)}{Precision(i)} + 1\right)}{TotalPatients}$$

Discussion

In this report we describe the application of machine learning models to predicting complications from gastric surgery using a dataset of 524 patients and 481 initial features. Our process included feature selection and the use of a Naïve Bayes classification model.

In order to facilitate our analysis, we used and developed SIH (Stay In Hospital) rate (or- Total Alert Rate, TAR), as a key evaluation method in this use case. SIH defines the percent of patients needed to be kept in the hospital for further monitoring, in order to achieve a specified recall for AL. When evaluating different models with the SIH metric, the recall was kept constant at 0.9.

We show that using pre surgery vanilla Naïve Bayes model (no feature selection) for a recall rate of 0.9, we attain the lowest SIH rate of 0.4 in comparison to other pre surgery models.

Conclusions

Each group of models can yield different insights on patients. The models that used Pre Surgery data can lead to inference that can affect decisions related to the surgical procedure. The models that used post surgery data can support decisions related to post-surgery treatment and management. We expect both approaches to be useful as support tools in clinical practice. When evaluating the best performing models, we will optimize for high rates of recall as these represent the number of expected ALs that we can predict in advance. Given a desired recall of 90%, the Pre Surgery model, considering all 53 features, yielded the lowest SIH rate for Pre Surgery data. For a different use case, given a desired recall of 90%, the feature disjointified model (DNB), considering 36 features, yields the lowest results for SIH rate for the Post Surgery data.

Limitations and Future Directions

When attempting to run the data through other methods, such as decision trees, we observed low performance. The observed test recall rates for these models where much lower than those obtained using Naïve Bayes. We therefore report results from Naïve Bayes only.

We note that the data is unbalanced since AL represents ~8% of the cases. This is in the nature of the diagnostic task we are addressing in this report. It is typically in the nature of many diagnostic tasks as the acute condition is often rare. Thus, we focus on recall as the key performance indicator, which is one aspect of how this imbalance is taken into account.

An additional limitation presented by the data was that the target class was binary. The models disregard the severity of the AL event, or the length of the hospital stay, for example. In future extensions of this work, this aspect can be considered.

Finally, more patient features should be considered such as medical historical data as in [10]. Bigger cohorts are also necessary for validation and in order to support higher quality inference with such data.

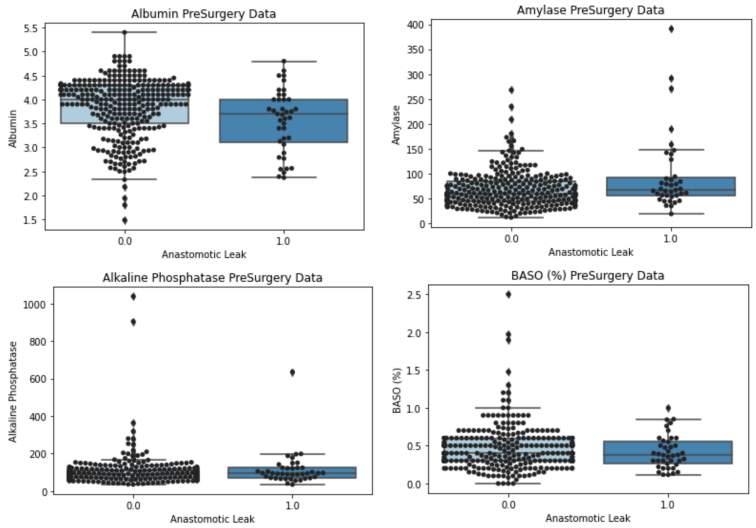
Acknowledgment

We thank all to all the team members that took part in this project. Both within Reichman University (IDC) and externally from the Sheba Hospital Staff. We thank the Yakhini Research Group for useful discussions.

Supplementary Information

1. Distribution of all data features

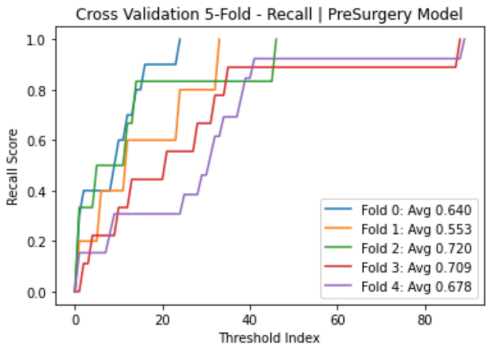
   Sample of data feature distribution is shown below. The data shown here is prior to any data preprocessing such as outlier detection and correction. Additional distributions available in code.



2. Link to excel with all data
   https://github.com/mayakerem/Anastomotic-Leak-Prediction/tree/main/data_files

3. 5-Fold cross validation results

   We report 5-fold cross validation results applied to Pre Surgery data model in order to preform cross validation due to small data set size.



4. Link to code

https://colab.research.google.com/drive/1-NG6rueQElxZjTg5i_pHkFLd3kBrVTgR?usp=sharing

References

[1] A. S. Rickles, J. C. Iannuzzi, K. N. Kelly, R. N. Cooney, D. A. Brown, M. Davidson, N. Hellenthal, C. Max, J. Johnson, J. DeTraglia, et al. AnastomoΘc leak or organ space surgical site infecΘon: what are we missing in our quality improvement programs? Surgery, 154(4):680–689, 2013.

[2] Amit, I., Iancu, O., Levy-Jurgenson, A. *et al.* CRISPECTOR provides accurate estimation of genome editing translocation and off-target activity from comparative NGS data. *Nat Commun* **12,** 3042 (2021). https://doi.org/10.1038/s41467-021-22417-4

[3] Bishop, C. M. (2016). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer. Section 8.1.

[4] A. S. Rickles, J. C. Iannuzzi, K. N. Kelly, R. N. Cooney, D. A. Brown, M. Davidson, N. Hellenthal,
C. Max, J. Johnson, J. DeTraglia, et al. Anastomotic leak or organ space surgical site infection: what are we missing in our quality improvement programs? Surgery, 154(4):680–689, 2013.

[5] Wilcoxon, Frank (Dec 1945). "Individual comparisons by ranking methods" (PDF). *Biometrics Bulletin*. 1 (6): 80–83. doi:10.2307/3001968. hdl:10338.dmlcz/135688. JSTOR 3001968.

[6] Graham J. G. Upton. “Fisher's Exact Test.” Journal of the Royal Statistical Society. Series A (Statistics in Society), vol. 155, no. 3, 1992, pp. 395–402. JSTOR, www.jstor.org/stable/2982890. Accessed 28 Aug. 2021.

[7] Yonatan Peleg, Shai Shefer, Leon Anavy, Alexandra Chudnovsky, Alvaro Israel, Alexander Golberg, Zohar Yakhini, Sparse NIR optimization method (SNIRO) to quantify analyte composition with visible (VIS)/near infrared (NIR) spectroscopy (350 nm-2500 nm), Analytica Chimica Acta, Volume 1051, 2019, Pages 32-40, ISSN 0003-2670, https://doi.org/10.1016/j.aca.2018.11.038.

[8] Benjamini Y, Hochberg Y (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing" (PDF). Journal of the Royal Statistical Society, Series B. 57 (1): 289–300. MR 1325392.

[9] Enerly E, Steinfeld I, Kleivi K, Leivonen SK, Aure MR, et al. (2011) miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors. PLOS ONE 6(2): e16915. https://doi.org/10.1371/journal.pone.0016915

[10] Karliczek, A., Harlaar, N.J., Zeebregts, C.J. et al. Surgeons lack predictive accuracy for anastomotic leakage in gastrointestinal surgery. Int J Colorectal Dis 24, 569–576 (2009). https://doi.org/10.1007/s00384-009-0658-6