

Machine Learning-Augmented Surveillance of Surgical Site Infections

Ugur Celik, Feifan Liu, Kimiyoshi J Kobayashi, Richard T Ellison 3rd, Yurima Guilarte-Walker, Deborah A Mack, Qiming Shi, Adrian Zai

Submitted to: JMIR Formative Research
on: April 01, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
---------------------------------	----------

Preprint
JMIR Publications

Machine Learning-Augmented Surveillance of Surgical Site Infections

Ugur Celik^{1*} MSc; Feifan Liu^{2*} PhD; Kimiyoshi J Kobayashi^{2, 3, 4*} MD, MBA; Richard T Ellison 3rd^{3, 5*} MD; Yurima Guilarte-Walker^{2*} MSc; Deborah A Mack^{5*} RN, CJS; Qiming Shi^{1*} MSc; Adrian Zai^{2*} MD, PhD, MPH

¹Center for Clinical and Translational Sciences University of Massachusetts Chan Medical School Worcester US

²Department of Population and Quantitative Health Sciences Division of Health Informatics and Implementation Science University of Massachusetts Chan Medical School Worcester US

³Department of Medicine University of Massachusetts Chan Medical School Worcester US

⁴UMass Memorial Medical Center Worcester US

⁵Infection Control Department UMass Memorial Medical Center Worcester US

*these authors contributed equally

Corresponding Author:

Ugur Celik MSc

Center for Clinical and Translational Sciences

University of Massachusetts Chan Medical School

55 Lake Avenue North

Worcester

US

Abstract

Background: Surgical Site Infection (SSI) is one of the most common healthcare-associated infections, comprising nearly 20% of all Healthcare-Associated Infections (HAI) in hospitalized patients. SSIs contribute to increased hospital length of stay, readmission, and healthcare costs, with an associated mortality rate twice that of non-infected patients.

Objective: To develop machine learning models to predict potential SSI cases following colon surgeries, thereby improving the efficiency of SSI surveillance and enabling healthcare professionals to better prioritize patient care.

Methods: Data was extracted from the EPIC Electronic Health Record system at a single academic center focusing on colon surgery patients between 2018 and 2023. The dataset included structured features such as demographics, medications, and vital signs, as well as unstructured clinical notes processed using natural language processing techniques. Logistic Regression, Random Forest, and XGBoost models were developed and trained to predict SSI risk. Cost-sensitive learning and Synthetic Minority Over-sampling Technique (SMOTE) were applied to handle the imbalanced dataset. Model evaluation was conducted using performance metrics including accuracy, precision, recall, and Area Under the Receiver Operating Characteristic curve (AUC-ROC).

Results: From a cohort of 1508 patients, 66 (4.4%) developed SSIs. The XGBoost model demonstrated the best overall performance with precision of 50%, recall of 38%, and an AUC-ROC score of 0.788. The Random Forest model achieved the highest precision (100%) but with lower recall (23%), while Logistic Regression showed higher recall (46%) but lower precision (10%). Patients who developed SSIs were significantly older (mean age 61.1 vs. 58.5 years), had higher ASA scores (78.8% vs. 45.3% with ASA score 3), more frequently had contaminated wounds (40.9% vs. 27.5%), and more commonly received steroids (83.3% vs. 58.7%).

Conclusions: Machine learning has shown promising results in enhancing the efficiency of SSI surveillance by enabling healthcare professionals to prioritize high-risk patients. While the models demonstrated strong overall accuracy, challenges related to imbalanced datasets remain. The XGBoost model provided the best balance between precision and recall, making it potentially the most clinically useful approach for SSI surveillance augmentation.

(JMIR Preprints 01/04/2025:75121)

DOI: <https://doi.org/10.2196/preprints.75121>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone.



Original Manuscript

Machine Learning-Augmented Surveillance of Surgical Site Infections

Authors: Ugur Celik,MS¹, Feifan Liu,PHD², Kimiyoshi J Kobayashi,MD,MBA^{2,3,4}, Richard T Ellison 3rd,MD^{3,4}, Yurima Guilarte-Walker,MS², Deborah Mack,RN,CJS³, Qiming Shi,MS¹, Adrian H Zai,MD,PHD,MPH²

Affiliations:

1. Center for Clinical and Translational Sciences, UMass Chan Medical School, Worcester, MA, USA.
2. Division of Health Informatics and Implementation Science, Department of Population and Quantitative Health Sciences, UMass Chan Medical School, Worcester, MA, USA.
3. UMass Memorial Medical Center, Worcester, MA, USA.
4. Department of Medicine, UMass Chan Medical School, Worcester, MA, USA

Corresponding Author:

Ugur Celik
Center for Clinical and Translational Sciences,
UMass Chan Medical School,
55 Lake Avenue North, Worcester, MA, USA.
Email: Ugur.Celik@umassmed.edu

Abstract

Background: Surgical Site Infection (SSI) is one of the most common healthcare-associated infections, comprising nearly 20% of all Healthcare-Associated Infections (HAI) in hospitalized patients. SSIs contribute to increased hospital length of stay, readmission, and healthcare costs, with an associated mortality rate twice that of non-infected patients.

Objective: To develop machine learning models to predict potential SSI cases following colon surgeries, thereby improving the efficiency of SSI surveillance and enabling healthcare professionals to better prioritize patient care.

Methods: Data was extracted from the EPIC Electronic Health Record system at a single academic center focusing on colon surgery patients between 2018 and 2023. The dataset included structured features such as demographics, medications, and vital signs, as well as unstructured clinical notes processed using natural language processing techniques. Logistic Regression, Random Forest, and XGBoost models were developed and trained to predict SSI risk. Cost-sensitive learning and Synthetic Minority Over-sampling Technique (SMOTE) were applied to handle the imbalanced

dataset. Model evaluation was conducted using performance metrics including accuracy, precision, recall, and Area Under the Receiver Operating Characteristic curve (AUC-ROC).

Results: From a cohort of 1508 patients, 66 (4.4%) developed SSIs. The XGBoost model demonstrated the best overall performance with precision of 50%, recall of 38%, and an AUC-ROC score of 0.788. The Random Forest model achieved the highest precision (100%) but with lower recall (23%), while Logistic Regression showed higher recall (46%) but lower precision (10%). Patients who developed SSIs were significantly older (mean age 61.1 vs. 58.5 years), had higher ASA scores (78.8% vs. 45.3% with ASA score 3), more frequently had contaminated wounds (40.9% vs. 27.5%), and more commonly received steroids (83.3% vs. 58.7%).

Conclusions: Machine learning has shown promising results in enhancing the efficiency of SSI surveillance by enabling healthcare professionals to prioritize high-risk patients. While the models demonstrated strong overall accuracy, challenges related to imbalanced datasets remain. The XGBoost model provided the best balance between precision and recall, making it potentially the most clinically useful approach for SSI surveillance augmentation.

Keywords: Surgical Site Infection; Machine Learning; healthcare-associated infections; surveillance; electronic health records; natural language processing; colon surgery; risk prediction; XGBoost; random forest

Introduction

Surgical Site Infections (SSIs) are a significant category of Healthcare-Associated Infections (HAIs), representing a serious challenge to healthcare systems globally. SSIs are estimated to account for nearly 20% of all HAIs among hospitalized patients [1]. In the United States, SSIs occur in approximately 2% to 4% of patients undergoing inpatient surgical procedures [2]. SSIs can lead to severe complications, including increased morbidity, extended hospital stays, higher readmission rates, and increased healthcare costs, ultimately impacting patient outcomes and putting a strain on healthcare resources [3-5]. The financial burden of SSIs is substantial, with associated costs reaching billions of dollars annually. SSIs are the third most costly type of healthcare-acquired infection, with an estimated cost of \$20,785 per patient case [6]. Patients with SSIs have a mortality rate twice that of non-infected patients [7].

The current surveillance process for SSIs is resource-intensive, requiring manual chart reviews by infection control practitioners to monitor surgical procedures and screen for potential infections. This manual surveillance is time-consuming, labor intensive, and detracts from direct patient care [8]. The Center for Disease Control and Prevention (CDC)'s National Healthcare Safety Network (NHSN) has established comprehensive guidelines to support the systematic tracking and identification of SSIs, yet the manual workload remains significant [9].

Several studies have applied machine learning (ML) and natural language processing (NLP) to improve the detection and prediction of SSIs. These approaches have shown promise in automating surveillance and improving accuracy by utilizing structured and unstructured clinical data. However, prior research has largely focused on retrospective detection rather than prospective prediction, relied

on limited data sources, or failed to integrate both structured and unstructured data [10-13].

To address these limitations, our study introduces an ML-based surveillance tool designed to enhance SSI monitoring. By integrating structured data and clinical notes, our approach improves detection efficiency and accuracy. Automating surveillance through ML reduces the burden on infection control practitioners, allowing them to focus on high-risk patients and critical tasks, ultimately improving patient outcomes and optimizing resource allocation.

Methods

Data Extraction and Preparation

Data for this study was extracted from the UMass Chan Data Lake, which is a copy of the EPIC Clarity EHR system at the hospital and refreshes weekly. This data repository comprises 2.5 million unique patients since November 2017. We focused on patients who underwent colon surgeries between 2018 and 2023, identified using Current Procedural Terminology (CPT) codes according to the National Healthcare Safety Network (NHSN) guidelines. Our cohort included 1508 procedures, with 66 confirmed SSI cases. The study received Institutional Review Board (IRB) exemption from UMass Chan Medical School as a quality improvement project.

To ensure clinical relevance, infection control nurses provided input during the term selection process for clinical note analysis and validated the model's outputs against real-world clinical scenarios. The dataset included 78 structured variables such as demographics, medications, laboratory results, and medical histories (Table 1). Additionally, two variables—SSI keyword count and SSI negation term count—were derived from unstructured clinical notes using NLP techniques. Specifically, we used the spaCy library [14] for tokenization and named entity recognition to identify relevant keywords associated with SSIs, such as “redness,” “swelling,” “drainage,” and “purulent discharge.” negation detection was performed using the NegEx algorithm [15], which allowed us to determine whether an identified keyword was negated in the context of the clinical note. For example, “no signs of infection” or “denies fever” would be identified as negated terms. These NLP tools enabled us to accurately quantify the occurrence of SSI-related terms and their context within the notes.

Table 1. Data Variables

Data Domain	Data Points
Demographics	Gender, Race, Ethnicity, Age, Smoking Status
Diagnosis	Obesity, Cancer, Diabetes, Immunological Disease, Depression, Dementia, Anemia, Heart Failure, AIDS, Alcohol

Encounter	Total Stay Days, Inpatient Stay
Lab	Hemoglobin, Culture, White Blood Cell, C-Reactive Protein
Medications	Antibiotics, Immunosuppressant, Steroid
Patient	Patient ID, Patient Encounter ID
Surgery	Surgery Class, Surgery Procedure Code, Physician ID, Number of Procedure, Surgery Duration, Surgery Recovery Duration, Department, Room Number, Wound Status, Anesthesia Type, ASA Score, Incision Closure, SSI
Vitals	BMI
NLP	SSI_Keyword, SSI_Negation
SSI Keywords	Fever, Nausea, Vomiting, Pain, Tenderness, Odynophagia, Dysphagia, hypotension, jaundice, dysuria, abscess, infection

Data Processing

Data preprocessing involved imputing missing values in numerical columns with column-wise means, one-hot encoding categorical variables, and validating data types for compatibility with machine learning algorithms. The dataset was processed using the Python pandas library [16] and then deployed to a secure workspace, the PLUM environment at UMass Chan, for further analysis [17].

Model Development

The dataset was split into training (80%) and validation (20%) sets using stratified sampling to maintain the class distribution. To address class imbalance, we applied cost-sensitive learning and the Synthetic Minority Over-sampling Technique (SMOTE) [19]. SMOTE works by creating synthetic examples of the minority class by interpolating between existing instances, thereby balancing the class distribution in the training set.

We developed three machine learning models: Logistic Regression [20], Random Forest [21], and XGBoost [22]. Each model was chosen based on specific strengths—Logistic Regression for interpretability, Random Forest for handling complex interactions, and XGBoost for efficient structured data analysis.

Logistic Regression was chosen due to its simplicity and interpretability, which allows healthcare

professionals to understand the relationships between features and outcomes, making it useful for clinical decision-making. Random Forest was selected for its ability to handle complex feature interactions and its robustness in managing missing data, making it effective for capturing non-linear patterns in the dataset. XGBoost was included for its high predictive performance and efficiency, particularly with structured datasets, and its ability to handle imbalanced data effectively through boosting techniques.

For Logistic Regression, we used L2 regularization [23] to prevent overfitting and applied polynomial features (adding interaction terms between features) to capture potential non-linear relationships in the data. For Random Forest, we optimized the number of trees, maximum depth, and minimum samples per leaf to ensure robustness. We used the same feature set for all models, including the polynomial features, to ensure a fair comparison. For XGBoost, we tuned hyperparameters including learning rate, maximum depth, and regularization parameters through grid search with cross-validation [24].

The structured data served as inputs for the models, while unstructured notes were processed with NLP techniques to identify potential SSI indicators.

Validation and Testing

Performance metrics, including accuracy, sensitivity (recall), specificity, precision, F1-score, and AUC-ROC, were calculated to comprehensively evaluate model performance, especially given the imbalanced nature of the dataset. Sensitivity and specificity help assess the model's ability to correctly identify positive and negative cases, respectively, while AUC-ROC offers an overall measure of discriminative performance. The final models were validated on the reserved 20% of the dataset to assess their generalizability.

Results

User Statistics

These are only examples of possible headings. Please feel free to use different headings to best describe your results.

Out of the 1508 patients in the study cohort, 66 (4.4%) developed SSIs. Table 2 presents the comprehensive characteristics of the study population, comparing patients with and without SSIs. The table is organized into demographic, clinical, surgical, and medication-related variables to facilitate interpretation.

Demographically, patients with SSIs were older (mean age 61.1 ± 8.9 years) compared to those without (mean age 58.5 ± 15.4 years), with most SSI cases (57.6%) occurring in the 61-80 age category. A higher proportion of females developed SSIs (65.2% vs. 53.1% in non-SSI group), though this difference did not reach statistical significance ($P=0.0717$).

The American Society of Anesthesiologists (ASA) score [25] differed significantly between groups ($P < 0.001$), with a substantially higher percentage of patients with SSIs having an ASA score of 3 (78.8% vs. 45.3%), indicating greater preoperative risk and comorbidity burden. Patients who developed SSIs showed a higher prevalence of several comorbidities, including diabetes (25.8% vs. 18.4%, $P=0.179$), depression (25.8% vs. 20.5%, $P=0.184$), anemia (36.4% vs. 29.3%, $P=0.277$), hypertension (54.5% vs. 45.4%, $P=0.184$), and chronic kidney disease (13.6% vs. 8.5%, $P=0.225$), though these differences individually did not reach statistical significance.

Surgical and procedural characteristics revealed important differences. Patients who developed SSIs underwent longer surgeries (mean duration 262.5 vs. 229.9 minutes), had a significantly higher number of procedures (mean 1.7 vs. 1.4, $P=0.0216$), and experienced substantially longer hospital stays (mean 8.1 vs. 6.3 days, $P=0.0004$). Wound classification showed significant variation, with contaminated wounds being more common in the SSI group (40.9% vs. 27.5%, $P=0.0261$) and clean-contaminated wounds being less common (36.4% vs. 55.4%, $P=0.0036$).

Laboratory and medication factors also demonstrated significant associations with SSI development. Abnormal white blood cell count were significantly more common in SSI patients (77.3% vs. 50.9%, $P<0.0001$), defined as WBC counts higher than 11 before surgery. Steroid use, categorized as receiving any steroid within 12 months prior to surgery, was dramatically higher in the SSI group (83.3% vs. 58.7%, $P=0.0001$). All patients who developed SSIs had received antibiotics.

Table 3 compares the performance metrics for each of the three machine learning models. The Logistic Regression model demonstrated an accuracy of 79%, with a precision of 10%, recall of 46%, F1-score of 0.16, and an AUC-ROC score of 0.706. The confusion matrix showed that out of 13 actual SSI cases in the test set, the model correctly identified 6 (true positives) but misclassified 7 (false negatives). Additionally, it incorrectly flagged 56 non-SSI cases as potential SSIs (false positives).

The Random Forest model achieved an accuracy of 97%, with a precision of 100%, recall of 23%, F1-score of 0.38, and an AUC-ROC score of 0.778. The confusion matrix revealed that the model correctly identified 3 out of 13 SSI cases (true positives) while missing 10 cases (false negatives). Notably, the model had no false positives, indicating high precision but lower sensitivity.

The XGBoost model showed a more balanced performance with an accuracy of 96%, precision of 50%, recall of 38%, F1-score of 0.43, and an AUC-ROC score of 0.788. The confusion matrix indicated that the model correctly identified 5 out of 13 SSI cases (true positives) while missing 8 cases (false negatives). It incorrectly flagged only 5 non-SSI cases as potential SSIs (false positives).

Among the three models, XGBoost demonstrated the highest AUC-ROC score (0.788) and F1-score (0.43), suggesting the best overall discriminative ability and balance between precision and recall. The Random Forest model showed perfect precision but lower recall, indicating it was highly conservative in predicting SSIs. Logistic Regression had the highest recall but the lowest precision, suggesting it was more liberal in flagging potential SSI cases but at the cost of many false positives.

Feature importance analysis revealed that the most predictive factors for SSI risk included ASA score, patient age, wound classification (particularly contaminated wounds), steroid use, and laboratory indicators such as white blood cell count. The NLP-derived features (SSI keyword count and negation term count) also contributed significantly to the models' predictive performance, highlighting the value of incorporating unstructured clinical notes in SSI risk assessment.

Table 2. Demographic and clinical characteristics by SSI status.

Variable	All Patients (n=1508)	No SSI (n=1442)	SSI (n=66)	P-value
Gender				0.0717
Female	808.0 (53.6%)	765 (53.1%)	43 (65.2%)	
Male	700.0 (46.4%)	677 (46.9%)	23 (34.8%)	
Race				0.5682
White	1323.0 (87.7%)	1262 (87.5%)	61 (92.4%)	
Black or African American	47.0 (3.1%)	45 (3.1%)	2 (3.0%)	
Asian	18.0 (1.2%)	16 (1.1%)	2 (3.0%)	
Other	120.0 (8.0%)	119 (8.3%)	1 (1.6%)	
Ethnicity				0.4252
Not Hispanic or Latino	1379.0 (91.4%)	1315 (91.2%)	64 (97.0%)	
Hispanic or Latino	114.0 (7.6%)	112 (7.8%)	2 (3.0%)	
Decline/Unknown	15.0 (1.0%)	15 (1.0%)	0 (0.0%)	
Age Category				0.0146
<18	15.0 (1.0%)	15 (1.0%)	0 (0.0%)	
18-40	145.0 (9.6%)	144 (10.0%)	1 (1.5%)	
41-60	605.0 (40.1%)	578 (40.3%)	27 (40.9%)	
61-80	653.0 (43.3%)	615 (42.9%)	38 (57.6%)	
>80	83.0 (5.5%)	83 (5.8%)	0 (0.0%)	
Age (years)	Mean: 58.6 ± 15.2	Mean: 58.5 ± 15.4	Mean: 61.1 ± 8.9	0.2589
BMI Category				0.0681
Underweight	47.0 (3.1%)	47 (3.3%)	0 (0.0%)	
Normal	426.0 (28.2%)	404 (28.4%)	22 (33.3%)	
Overweight	509.0 (33.8%)	488 (34.3%)	21 (31.8%)	
Obese I	306.0 (20.3%)	295 (20.7%)	11 (16.7%)	
Obese II	150.0 (9.9%)	138 (9.7%)	12 (18.2%)	
Obese III	52.0 (3.4%)	52 (3.7%)	0 (0.0%)	
BMI (kg/m ²)	Mean: 28.4 ± 6.8	Mean: 28.4 ± 6.9	Mean: 28.9 ± 5.8	0.5055
ASA Score				0.0000

1	14.0 (0.9%)	14 (1.0%)	0 (0.0%)	
2	756.0 (50.1%)	743 (51.5%)	13 (19.7%)	
3	705.0 (46.8%)	653 (45.3%)	52 (78.8%)	
4-5	33.0 (2.2%)	32 (2.2%)	1 (1.5%)	
Comorbidities				
Number of Comorbidities	Mean: 1.9 ± 1.7	Mean: 1.9 ± 1.7	Mean: 2.2 ± 1.8	0.0961
Obesity	325.0 (21.6%)	308 (21.4%)	17 (25.8%)	0.4860
Cancer	287.0 (19.0%)	276 (19.1%)	11 (16.7%)	0.7337
Diabetes	282.0 (18.7%)	265 (18.4%)	17 (25.8%)	0.1795
Depression	313.0 (20.8%)	296 (20.5%)	17 (25.8%)	0.3846
Anemia	447.0 (29.6%)	423 (29.3%)	24 (36.4%)	0.2779
Heart Failure	59.0 (3.9%)	57 (4.0%)	2 (3.0%)	1.0000
Hypertension	691.0 (45.8%)	655 (45.4%)	36 (54.5%)	0.1841
Chronic Kidney Disease	132.0 (8.8%)	123 (8.5%)	9 (13.6%)	0.2252
Chronic Liver Disease	5.0 (0.3%)	4 (0.3%)	1 (1.5%)	0.2007
Surgical Characteristics				
Surgery Duration (minutes)	Mean: 231.3 ± 105.5	Mean: 229.9 ± 103.6	Mean: 262.5 ± 137.3	0.1400
Number of Procedures	Mean: 1.4 ± 0.8	Mean: 1.4 ± 0.7	Mean: 1.7 ± 1.1	0.0216
Length of Stay (days)	Mean: 6.4 ± 10.4	Mean: 6.3 ± 10.5	Mean: 8.1 ± 6.8	0.0004
Inpatient Surgery	231.0 (15.3%)	220 (15.3%)	11 (16.7%)	0.8916
Wound Classification				
Clean Wound	121.0 (8.0%)	115 (8.0%)	6 (9.1%)	0.9246
Clean-Contaminated Wound	823.0 (54.6%)	799 (55.4%)	24 (36.4%)	0.0036
Contaminated Wound	424.0 (28.1%)	397 (27.5%)	27 (40.9%)	0.0261
Anesthesia				
General Anesthesia	1450.0 (96.2%)	1384 (96.0%)	66 (100.0%)	0.1763
General + Epidural	54.0 (3.6%)	54 (3.7%)	0 (0.0%)	0.1685
Laboratory Results				
Hemoglobin Flag	1503.0 (99.7%)	1437 (99.7%)	66 (100.0%)	1.0000
C-Reactive Protein Flag	237.0 (15.7%)	230 (16.0%)	7 (10.6%)	0.3204
White Blood Cell Flag	785.0 (52.1%)	734 (50.9%)	51 (77.3%)	0.0000
Positive Culture	8.0 (0.5%)	8 (0.6%)	0 (0.0%)	1.0000
Medications				
Antibiotic Use	1467.0 (97.3%)	1401 (97.2%)	66 (100.0%)	0.2563
Immunosuppressive Therapy	73.0 (4.8%)	70 (4.9%)	3 (4.5%)	1.0000

Steroid Use	901.0 (59.7%)	846 (58.7%)	55 (83.3%)	0.0001
-------------	---------------	-------------	------------	--------

*Not all variables are shown for brevity.

*P-values calculated using Chi-square or Fisher's exact test for categorical variables and t-test for continuous variables.

*ASA Score: American Society of Anesthesiologists physical status classification system.

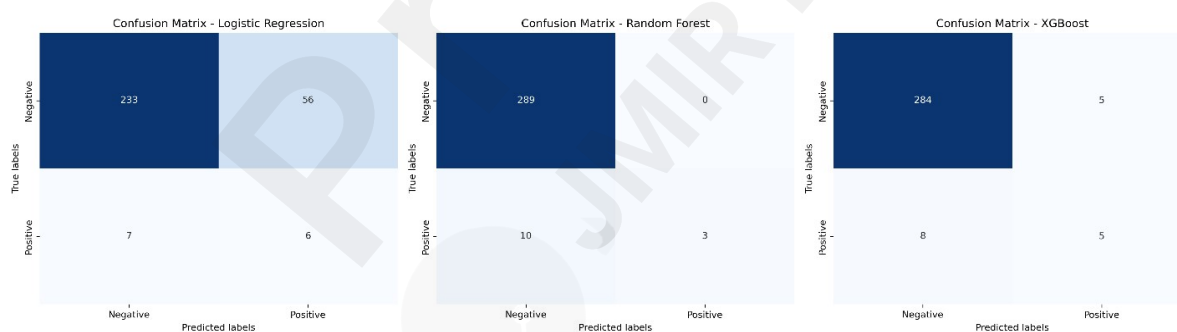
Table 3. Performance metrics across machine learning models.

Model	Precision (%)	Recall (%)	F1-Score	AUC-ROC
Logistic Regression	10	46	0.16	0.706
Random Forest	100	23	0.38	0.778
XGBoost	50	38	0.43	0.788

*Metrics calculated on a validation set of 302 patients (289 non-SSI, 13 SSI).

*AUC-ROC: Area under the receiver operating characteristic curve.

Figure 1. Confusion Matrices for each model



Discussion

Principal Results

All models demonstrated high overall accuracy, but their performance in identifying SSI cases varied considerably. The XGBoost model provided the best balance between precision and recall, making it

potentially the most clinically useful approach for SSI surveillance augmentation. The Random Forest model's perfect precision but low recall suggests it could be valuable in contexts where false positives must be minimized, though at the cost of missing many true cases. Conversely, the Logistic Regression model's higher recall but lower precision might be preferable in scenarios where identifying as many potential SSI cases as possible is the priority, even if it generates more false positives.

The differences in model performance highlight the trade-offs inherent in predictive modeling for rare events like SSIs. High accuracy can be misleading in imbalanced datasets, as simply predicting all cases as negative would still yield high accuracy due to the predominance of non-SSI cases. This underscores the importance of evaluating models using metrics beyond accuracy, particularly precision, recall, and AUC-ROC.

The most important predictive factors identified by our models align with known clinical risk factors for SSIs, including older age, higher ASA scores, contaminated wounds, and certain laboratory abnormalities. This concordance with clinical knowledge enhances the interpretability and credibility of our models.

The integration of NLP-derived features from clinical notes represents a novel aspect of our approach. By quantifying the occurrence of SSI-related terms and accounting for negations in clinical documentation, we were able to capture valuable information that might not be reflected in structured data alone. This highlights the potential of leveraging unstructured data in healthcare analytics to enhance predictive performance.

Limitations

Our evaluation showed that SMOTE was more effective than cost-sensitive learning in improving model performance for the minority class, increasing recall by the approximately 15% while maintaining similar precision. However, both approaches still exhibited some bias toward the majority class, particularly the Random Forest model. The severe class imbalance in our dataset (only 4.4% SSI cases) presents a significant challenge for model development and evaluation. Although we employed techniques to address this imbalance, the models' performance in identifying positive cases remained suboptimal, as evidenced by the modest recall values.

Additionally, the use of retrospective data from a single healthcare center may introduce biases related to specific patient populations and clinical practices, potentially limiting the generalizability of the results. The predictive factors identified in our specific setting might not hold the same importance in other healthcare environments with different patient demographics, surgical practices, or infection control protocols.

Furthermore, our approach to processing unstructured clinical notes, was limited to keyword counting and negation detection. More sophisticated NLP techniques such as embedding-based methods or transformer models might yield better feature extraction and ultimately enhance

predictive performance [26-28].

Comparison with Prior Work

Previous studies have explored various approaches to SSI prediction using machine learning. Thirukumaran et al. [29] applied NLP to clinical notes and achieved a sensitivity and positive predictive value of 97% in identifying SSIs, which exceeds the performance of our models. However, their approach focused on detection rather than prediction, analyzing notes from the post-infection period. Our study distinguishes itself by using only data available prior to infection onset, making it more suitable for prospective risk assessment.

Mandagani et al. [30] employed logistic regression, decision trees, and boosted decision trees using preoperative blood test results, achieving an AUC of 86%, which is higher than our best performing model (AUC-ROC of 0.788 for XGBoost). However, their study utilized a more balanced dataset and focused specifically on blood test parameters, whereas our approach integrated a broader range of clinical variables.

Shen et al. [31] used automated lexicon generation from clinical narratives, which is similar to our NLP approach but more sophisticated in its linguistic analysis. Da Silva et al. [32] utilized text mining and ML methods on surgical descriptions and postoperative records, achieving an AUC-ROC of approximately 80%, which is comparable to our XGBoost model performance.

The key advantage of our approach is the combination of structured clinical data with NLP-processed unstructured notes, providing a more comprehensive view of patient risk factors. Additionally, our focus on predicting rather than detecting SSIs makes our model more valuable for proactive clinical intervention.

Conclusions

This study demonstrates that ML models can effectively assist in the surveillance of surgical site infections, helping healthcare professionals prioritize patient care. The key contributions of this study include the development of an ML-based tool that leverages both structured and unstructured clinical data, providing an innovative solution to improve SSI monitoring efficiency.

The XGBoost model emerged as the most promising approach, offering a good balance between precision and recall while maintaining high overall accuracy. While none of the models achieved both high precision and high recall simultaneously, they each offer distinct advantages that could be valuable in different clinical contexts.

The model aids the identification process, thereby reducing the administrative burden on healthcare

staff, while also providing timely, data-driven insights to enhance clinical outcomes. By triaging patients based on predicted risk, the model can help infection control practitioners allocate their time and attention more efficiently, focusing on cases with the highest likelihood of SSI.

Future research should prioritize incorporating advanced techniques to handle data imbalance, such as combining resampling strategies with cost-sensitive learning, and expanding the dataset to include diverse healthcare settings to improve generalizability. Exploring ensemble approaches [33] that combine the strengths of different models could also enhance predictive performance. Additionally, more sophisticated NLP techniques could be employed to extract richer features from clinical notes, potentially improving the models' ability to capture subtle indicators of infection risk. Dataset augmentation methods using synthetic data generation [34] could also help address the class imbalance issues that continue to challenge predictive modeling in this domain.

Acknowledgements

We thank the Infection Prevention and Control Staff at UMass Memorial Medical Center (Deborah A. Mack, Melissa Dabkowski, Maria Del Pilar Diaz Padro) for their invaluable support and collaboration, as well as the UMass Chan Medical School IT Department for their essential technical assistance. Their contributions were vital to the success of this project. The grant is funded by The Doctors Company Foundation and the UMCCTS Pilot Project Program by Grant #UL1-TR001453.

Conflicts of Interest

None declared.

Abbreviations

ASA:	American	Society	of	Anesthesiologists
AUC:	area	under	the	curve
AUC-ROC:	area	under	the receiver operating	characteristic curve
CDC:	Centers	for	Disease Control	and Prevention
CPT:	Current	Procedural		Terminology
EHR:	electronic	health		record
HAI:	healthcare-associated			infection
ML:	machine			learning
NHSN:	National	Healthcare	Safety	Network
NLP:	natural	language		processing
SMOTE:	Synthetic	Minority	Over-sampling	Technique
SSI:	surgical site infection			

Multimedia

Appendix

1

Detailed model performance metrics including confusion matrices and ROC Curve plots for each machine learning model.

Figure 1. Confusion Matrices for each model

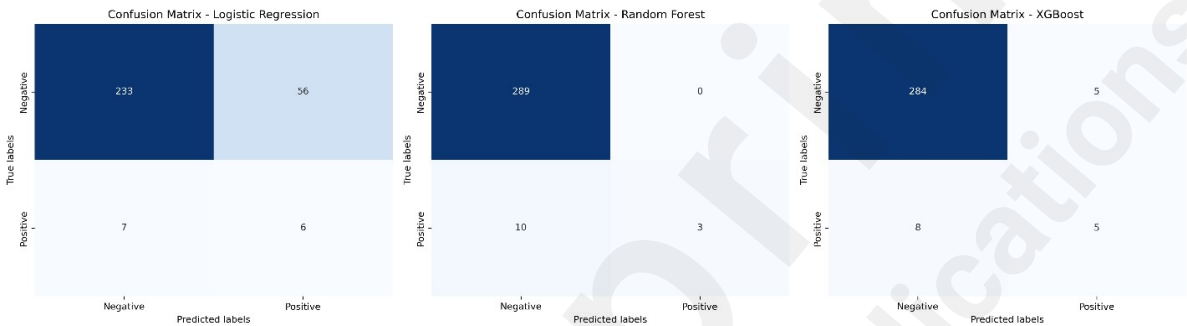
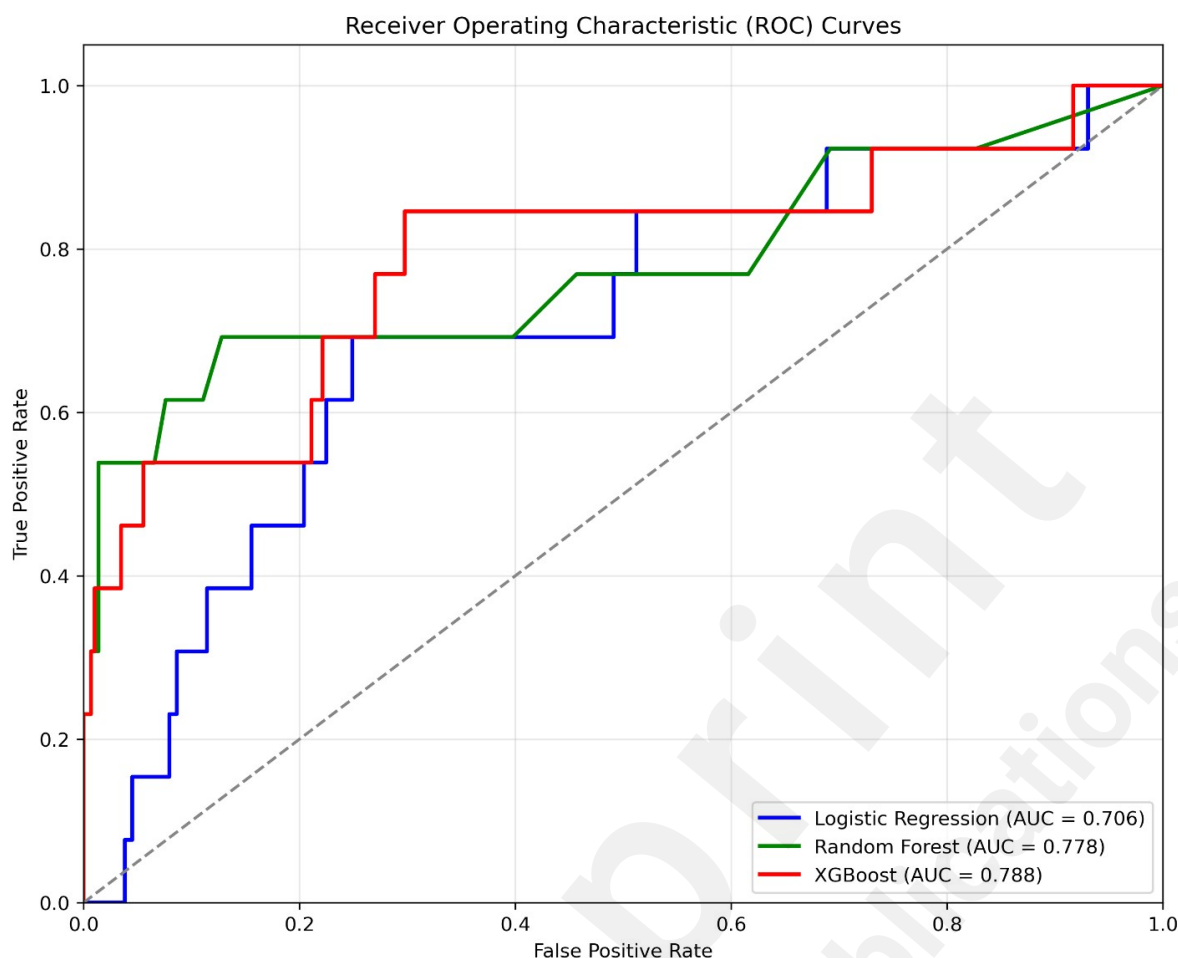


Figure 2. Receiver Operating Characteristics(ROC) Curve Comparison for each model.



References

1. JB Battles, SL Farr, DA Weinberg. From research to nationwide implementation: the impact of AHRQ's HAI prevention program. *Med Care* 2014;52(2 Suppl 1):91-96. PMID:24430269
2. Awad SS. Adherence to surgical care improvement project measures and post-operative surgical site infections. *Surg Infect (Larchmt)* 2012;13(4):234-237. PMID:22913334
3. RP Merkow, MH Ju, JW Chung, BL Hall, ME Cohen, MV Williams, TC Tsai, CY Ko, KY Bilimoria. Underlying reasons associated with hospital readmission following surgery in the United States. *JAMA* 2015;313(5):483-495. PMID:25647204
4. de Lissovoy G, Fraeman K, Hutchins V, Murphy D, Song D, Vaughn BB. Surgical site infection: incidence and impact on hospital utilization and treatment costs. *Am J Infect Control* 2009;37(5):387-397. PMID:19398246
5. Sparling KW, Ryckman FC, Schoettker PJ, Byczkowski TL, Helpling A, Mandel K, Panchanathan A, Kotagal UR. Financial impact of failing to prevent surgical site infections. *Qual Manag Health Care* 2007;16(3):219-225. PMID:17627217
6. Magill SS, Edwards JR, Bamberg W, Beldavs ZG, Dumyati G, Kainer MA, Lynfield R, Maloney M, McAllister-Hollod L, Nadle J, Ray SM, Thompson DL, Wilson LE, Fridkin SK. Multistate point-prevalence survey of health care-associated infections. *N Engl J Med* 2014;370(13):1198-1208. PMID:24670166
7. Anderson DJ, Podgorny K, Berríos-Torres SI, et al. Strategies to prevent surgical site infections in acute care hospitals: 2014 update. *Infect Control Hosp Epidemiol.*

2014;35(6):605-627.

PMID:24799638

8. Mangram A, Horan T, Pearson M, Silver L, Jarvis W. Guideline for Prevention of Surgical Site Infection, 1999. *Infection Control & Hospital Epidemiology* 1999;20(4):247-280. PMID:10219875
9. Centers for Disease Control and Prevention. Surgical Site Infection (SSI) Event. CDC; 2022. URL: <https://www.cdc.gov/nhsn/pdfs/pscmanual/9pscscsscurrent.pdf> [accessed 2024-01-15]
10. Sanger PC, van Ramshorst GH, Mercan E, et al. A Prognostic Model of Surgical Site Infection Using Daily Clinical Wound Assessment. *J Am Coll Surg*. 2016;223(2):259-270.e2. PMID:27212006
11. Meeks DW, Lally KP, Carrick MM, et al. Compliance with guidelines to prevent surgical site infections: As simple as 1-2-3? *Am J Surg*. 2011;201(1):76-83. PMID:20573335
12. Ridgeway S, Wilson J, Charlet A, Kafatos G, Pearson A, Coello R. Infection of the surgical site after arthroplasty of the hip. *J Bone Joint Surg Br*. 2005;87(6):844-850. PMID:15911671
13. Koek MBG, Hopmans TEM, Soetens LC, et al. Adhering to a national surgical care bundle reduces the risk of surgical site infections. *PLoS One*. 2017;12(9):e0184200. PMID:28880903
14. Honnibal M, Montani I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017.
15. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform*. 2001;34(5):301-310. PMID:12123149
16. McKinney W. Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*. 2010:56-61.
17. Zai AH, Wong S, Guilarte-Walker Y, Langlois P, Coleman B, Soni A, McManus DD, Luzuriaga K. Implementing virtual desktops for clinical research at an academic health center: a case report. *JAMIA Open* 2024;7(3):ooae083. PMID:39206281
18. Elkan C. The foundations of cost-sensitive learning. In: *International Joint Conference on Artificial Intelligence*. 2001:973-978.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002;16:321-357.
20. Hosmer DW, Lemeshow S, Sturdivant RX. *Applied Logistic Regression*. 3rd ed. Wiley; 2013.
21. Breiman L. Random Forests. *Machine Learning*. 2001;45:5-32.
22. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016:785-794.
23. Ng AY. Feature selection, L1 vs. L2 regularization, and rotational invariance. In: *Proceedings of the 21st International Conference on Machine Learning*. 2004:78.
24. Bergstra J, Bengio Y. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*. 2012;13:281-305.
25. Doyle DJ, Garmon EH. American Society of Anesthesiologists Classification (ASA Class). In: *StatPearls. Treasure Island (FL): StatPearls Publishing; 2022*.
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional

Transformers for Language Understanding. In: Proceedings of NAACL-HLT. 2019:4171-4186.

27. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240. PMID:31501885
28. Alsentzer E, Murphy JR, Boag W, et al. Publicly Available Clinical BERT Embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019:72-78.
29. Thirukumaran CP, Zaman A, Rubery PT, Calabria C, Li Y, Ricciardi BF, Bakhsh WR, Kautz H. Natural Language Processing for the Identification of Surgical Site Infections in Orthopaedics. *J Bone Joint Surg Am* 2019;101(24):2167-2174. PMID:31800431
30. Mandagani P, Coleman S, Zahid A, Pugel Ehlers A, Basu Roy S, De Cock M. Machine Learning Models for Surgical Site Infection Prediction. *AMIA KDDM-WG Symposium* 2016.
31. Shen F, Larson DW, Naessens JM, Habermann EB, Liu H, Dimango EP. Detection of Surgical Site Infection Utilizing Automated Feature Generation in Clinical Notes. *Journal of Healthcare Informatics Research* 2018;3(3):267-282. doi:10.1007/s41666-018-0042-9
32. da Silva DA, ten Caten CS, dos Santos RP, Fogliatto FS, Hsuan J. Predicting the occurrence of surgical site infections using text mining and machine learning. *PLoS ONE* 2019;14(12):e0226272. PMID:31881032
33. Dietterich TG. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer; 2000:1-15.
34. Park N, Mohammadi M, Gorde K, Jaeger S, Chen H, Liang X. Data synthesis based on generative adversarial networks. In: Proceedings of the VLDB Endowment. 2018;11(10):1071-1083.