

Original Research

Finite-Sample Instability, Threshold Sensitivity, and Error Propagation in Early Anastomotic Leak Prediction

Hakan Celik¹, Burak Sahin² and Tugrul Arslan³

¹Bandirma Onyedi Eylul University, Department of Computer Engineering, Sehit Astsabay Mustafa Soner Varlik Campus, 10200 Bandirma, Turkey.

²Hitit University, Department of Health Informatics, Cevre Yolu Bulvari No: 8, 19030 Corum, Turkey.

³Duzce University, Department of Biomedical Engineering, Konuralp Campus, 81620 Duzce, Turkey.

Abstract

Clinical prediction studies for postoperative complications often summarize performance by a small set of point estimates, even when the event is uncommon and the positive class in the test set is numerically small. In that setting, the apparent ranking of algorithms can be more fragile than it first appears. This paper presents an empirical uncertainty audit of early anastomotic leak prediction from structured perioperative records. The analysis reinterprets a reported held-out experiment not as a fixed leaderboard, but as a finite-resolution sample in which recall, precision, alert burden, and threshold choice are all subject to nontrivial instability. The paper examines three empirical questions. First, how much statistical resolution is actually available when only a small number of positive cases determine sensitivity estimates and thresholded comparisons? Second, how sharply can an F1-selected operating point distinguish between neighboring model families when one or two admissions can materially alter the observed metric balance? Third, how do these uncertainties scale when the model outputs are translated into service-level decisions such as alerts, workups, or intensified surveillance? The results indicate that the informative content of the experiment lies less in a simple declaration of the best model than in the geometry of uncertainty around the reported operating points. The stronger nonlinear configurations remain clinically interesting, but the observed gaps between several of them are narrower than their point estimates imply. At the same time, the differences in false-positive accumulation are large enough to matter under plausible error costs. The paper argues that rare-event surgical modeling should report interval structure, threshold volatility, and cost-ratio switching behavior alongside standard discrimination metrics, because these quantities better describe what the experiment actually establishes.

1. Introduction

Predictive modeling for surgical complications is usually narrated as a race among algorithms [1]. A development cohort is assembled, several model classes are trained, a held-out set is reserved, and the resulting scores are compared through a few familiar metrics. This workflow is technically reasonable, but its interpretive habits are often too coarse for uncommon outcomes. When the endpoint is rare, the evaluation sample is numerically asymmetric from the beginning. Most admissions in the test set are controls, only a small minority are cases, and thresholded quantities such as recall and precision are therefore carried by a relatively small number of positive observations. This means that the apparent certainty of a performance table can exceed the true resolution of the experiment. An observed advantage may reflect robust structure in the data, but it may also be partly a consequence of threshold choice, finite-sample discreteness, and the fact that one or two event assignments can move a clinically emphasized metric by several percentage points.

Anastomotic leak after colorectal surgery is a useful setting in which to examine this issue. The complication is serious enough that even moderate predictive signal attracts attention, but infrequent enough that every evaluation exercise is exposed to class imbalance. In such a setting, it is easy to overread

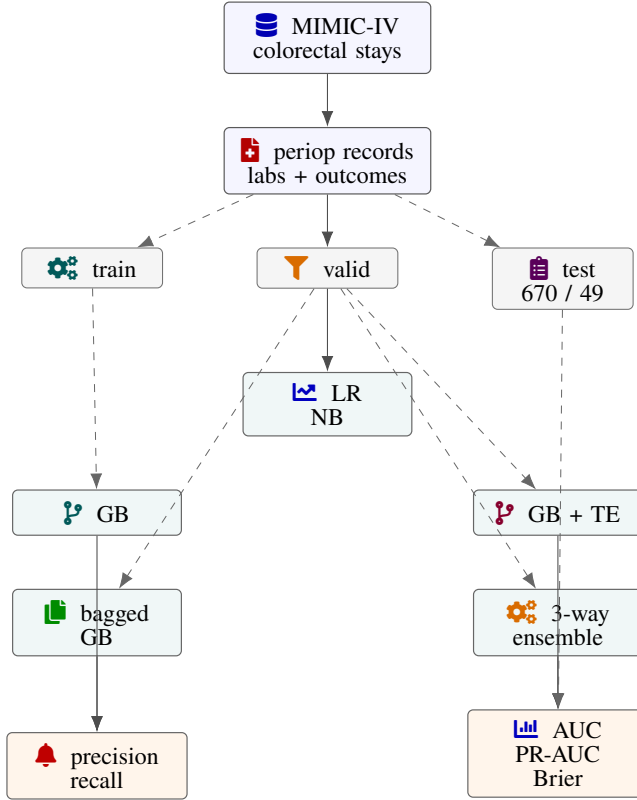


Figure 1: Empirical frame of the leak-prediction study. Structured perioperative records are split into development and held-out evaluation partitions, then compared across linear, probabilistic, boosting, bagged, and ensemble configurations. The figure emphasizes that the experiment is a thresholded comparison under class imbalance rather than a simple leaderboard, so discrimination, calibration, and alert burden must be read together.

point estimates. A model with recall 0.49 may appear clearly preferable to a model with recall 0.41, yet if the test set contains fewer than fifty positive admissions, the observed difference corresponds to only a handful of patients. Similarly, a model with precision 0.43 may appear meaningfully more selective than one with precision 0.34, but the width of any reasonable interval estimate can be substantial when the number of predicted-positive admissions is itself small. None of this implies that the results are uninformative. It means that the experiment is being conducted at a finite statistical resolution that should be made explicit.

This is especially important when the operating threshold is selected by a criterion such as F_1 . The threshold maximizing F_1 is not a biologically privileged point [2]. It is a numerical compromise between false negatives and false positives under a specific validation sample. When the positive class is small and the score distributions of neighboring models overlap, the chosen threshold can be unstable. Small score perturbations can reorder the top few admissions near the cutpoint, and because each positive event has disproportionate leverage in a low-prevalence setting, the resulting thresholded metrics can change in coarse steps. A model may therefore appear to achieve a finely tuned precision-recall balance even though the observed point lies on a noisy, locally jagged surface.

A second problem follows from the way rare-event models are discussed in clinical settings. Surgeons, intensivists, and quality teams rarely interact with abstract discrimination. They interact with alerts, escalations, workups, missed cases, and false alarms. The operational effect of a model depends not only on its ranking ability but also on how many patients are placed above threshold and how uncertain that thresholded volume is. A test set of modest size can support a reasonable estimate of the direction

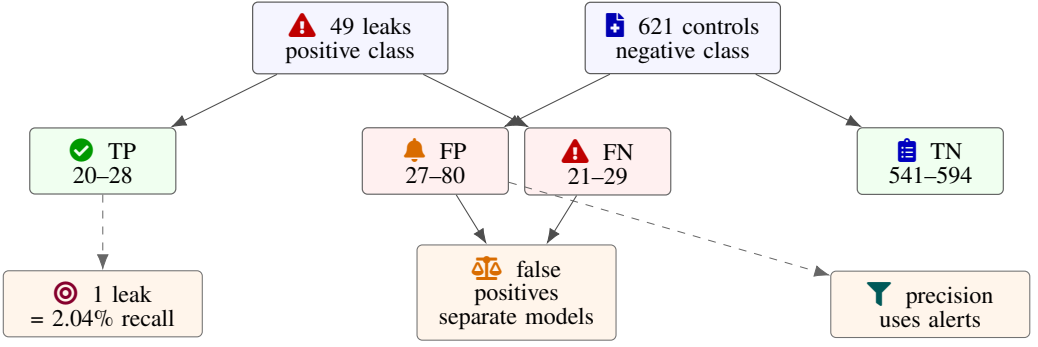


Figure 2: Finite-sample confusion geometry of the held-out cohort. Because only 49 admissions are positive, recall is intrinsically coarse: one retrieved leak shifts sensitivity by about 2.04 percentage points. By contrast, much of the practical separation among the stronger models is driven by false-positive accumulation, which changes the size of the predicted-positive set and therefore the stability of precision and downstream workup volume.

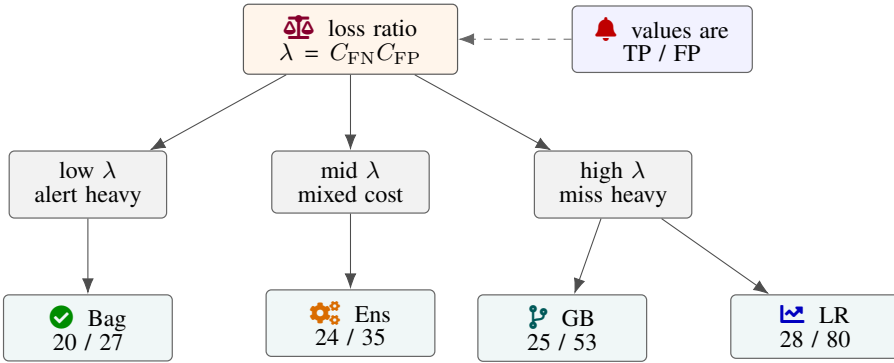


Figure 3: Decision-theoretic switching view of model choice. Approximate held-out true-positive and false-positive counts are shown as compact TP/FP summaries. When false-positive cost dominates, narrow-alert models such as bagged boosting become attractive; in intermediate regimes the ensemble can be preferred; and when missed leaks are penalized more heavily, higher-recall operating modes such as logistic regression or plain boosting can become competitive despite larger alert burden.

of model behavior, but it may support a much less precise estimate of downstream alert burden once the model is carried into larger service volumes. For this reason, a proper interpretation of the experiment should include not only what happened in the held-out sample, but also how sampling error propagates when those rates are projected to a realistic surgical population.

A third issue concerns model comparison itself. It is tempting to identify the single best method and stop there. Yet in clinical prediction, model choice often depends on the relative harm of false negatives and false positives rather than on a universal scalar score. If a missed leak is treated as much more costly than an unnecessary workup, a higher-recall model may be preferred despite lower precision. If the clinical environment is already saturated with low-yield alerts, the opposite may hold. What matters empirically is the switching structure: at what false-negative-to-false-positive cost ratio does one model become preferable to another? That question is almost never addressed directly, even though the information required to approximate it is frequently available.

The present paper addresses these issues through a secondary empirical analysis of a reported early postoperative leak prediction experiment [3]. The goal is not to re-engineer the underlying pipeline or to introduce a new algorithmic family. The goal is to interrogate what the existing experiment can actually tell us once one accounts for finite positive counts, coarse threshold resolution, and explicit error costs.

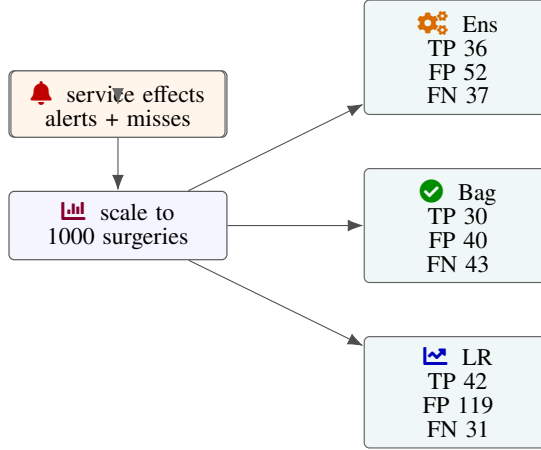


Figure 4: Error propagation from the held-out experiment to service-level throughput. The boxes show approximate counts per 1000 surgical admissions implied by the reported operating points: bagged boosting yields fewer alerts but more missed leaks, the ensemble occupies a middle regime, and logistic regression retrieves more positives at the price of substantially higher false-positive accumulation. This makes clear why point estimates alone are insufficient for clinical planning: alert burden and missed-case volume scale differently across models.

This is still an empirical paper in the strict sense. Every substantive calculation is tied to a reported cohort size, event count, or model metric. What changes is the object of study. Instead of asking which algorithm occupies the top line of a performance table, the paper asks how stable that ordering is, how much uncertainty surrounds the thresholded metrics, and how the observed rates would expand into service-level consequences under different decision assumptions.

The analysis proceeds in several steps. First, the data-generating frame of the published experiment is formalized, and the reported summary metrics are translated into approximate confusion counts. Second, standard finite-sample uncertainty calculations are used to show how wide the plausible ranges are for recall, precision, and AUC when the test set contains only a limited number of events. Third, the threshold selected by F_1 is treated as a noisy estimator, and the local sensitivity of the metric surface to one-case perturbations is examined. Fourth, a decision-theoretic loss family is introduced so that pairwise switching boundaries among the models can be stated explicitly rather than left implicit. Fifth, the uncertainty is propagated from the held-out set to larger service volumes, making visible the range of likely alert counts and missed cases per unit throughput. Finally, the leading predictors and the data-quality description are interpreted as an observation architecture that helps explain why certain model classes perform as they do.

The paper remains deliberately restrained. It does not claim that finite-sample instability nullifies the original modeling result. Nor does it claim that interval overlap implies that all models are equivalent. Instead, it argues that low-prevalence surgical prediction should be read through a sharper statistical lens than is customary [4]. Point estimates remain useful, but they are not the whole result. The uncertainty surrounding them, the discrete sensitivity of threshold selection, and the cost-ratio regimes in which model preference flips are all part of what the experiment has established. In a rare-event setting, that broader view is not an optional refinement. It is the difference between a leaderboard reading and an actual inferential reading.

2. Empirical Frame

The empirical basis for the present analysis is the postoperative leak prediction experiment reported by Sadanandan (2024) [5], which identified 3,331 colorectal resection admissions from MIMIC-IV, observed an anastomotic leak rate of 7.6%, evaluated six model configurations on a held-out test set of

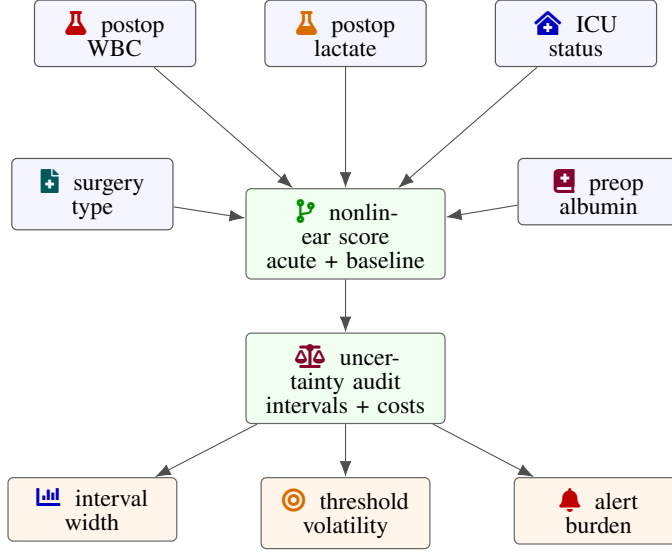


Figure 5: Observation architecture behind the uncertainty audit. Acute postoperative physiology, care-intensity indicators, operative context, and baseline reserve variables feed a nonlinear scoring layer, after which the inferential focus shifts from raw leaderboard ranking to interval structure, threshold sensitivity, and alert consequences. This view links the reported top predictors to the paper’s broader claim that the experiment is most informative when its uncertainty geometry is made explicit.

670 admissions with 49 positive outcomes, and reported AUC, PR-AUC, Brier score, precision, recall, and ranked predictor importance for each model. The models were logistic regression, naive Bayes, gradient boosting, gradient boosting with target encoding, bagged gradient boosting using five base learners, and a three-way ensemble. The top predictive variables in the gradient boosting analysis included postoperative white blood cell count, intensive care unit admission status, postoperative lactate, surgery type, and preoperative albumin, while the data-quality description noted low missingness for routine perioperative laboratory values and higher missingness for features tied to preoperative or outpatient documentation.

Those reported quantities are sufficient to reconstruct the approximate thresholded behavior of the six classifiers. Let $n = 670$ denote the held-out test size, $n_1 = 49$ the number of leak cases, and $n_0 = 621$ the number of non-leak admissions. For model m , let p_m be the reported precision and r_m the reported recall at the threshold selected on the validation set by F_1 . Then the implied confusion geometry is

$$\begin{aligned}
 \widehat{TP}_m &= r_m n_1 \\
 \widehat{FN}_m &= n_1 - \widehat{TP}_m \\
 \widehat{FP}_m &= \widehat{TP}_m \left(\frac{1 - p_m}{p_m} \right) \\
 \widehat{TN}_m &= n_0 - \widehat{FP}_m.
 \end{aligned} \tag{2.1}$$

Because the published precision and recall values were rounded to two decimal places, the counts reconstructed from them are themselves approximate. That does not materially limit the present analysis, since the purpose is not to recover exact patient-level labels but to examine the scale on which the reported thresholds operated. In a test set with only forty-nine positive admissions, the difference between a model retrieving twenty and one retrieving twenty-four positives is already meaningful; the uncertainty introduced by rounding to two decimals is negligible relative to that scale.

The reconstructed counts are informative. Logistic regression corresponds to roughly 27.9 true positives, 21.1 false negatives, 79.5 false positives, and 541.5 true negatives. Naive Bayes corresponds to approximately 34.8 true positives and 255.1 false positives [6]. Gradient boosting corresponds to roughly 25.0 true positives and 53.1 false positives. The target-encoded variant yields about 21.1 true positives and 40.9 false positives. Bagged gradient boosting yields approximately 20.1 true positives and 26.6 false positives. The three-way ensemble yields about 24.0 true positives and 34.6 false positives. Even before interval estimation begins, these approximations already reveal that much of the model differentiation lies in false-positive accumulation rather than in the number of retrieved positives.

Several additional thresholded quantities follow immediately. Let $k_m = \widehat{TP}_m / \widehat{FP}_m$ denote the number of admissions predicted positive by model m , and let $\alpha_m = k_m n$ be the selected fraction of the test set. Then α_m is roughly 0.160 for logistic regression, 0.433 for naive Bayes, 0.117 for gradient boosting, 0.093 for target-encoded gradient boosting, 0.070 for bagged gradient boosting, and 0.087 for the three-way ensemble. This alone has a clear operational meaning. The bagged model placed only about 7.0% of the test cohort above threshold, while logistic regression placed 16.0% above threshold and naive Bayes placed more than 43% above threshold. The principal experimental contrast is therefore not simply whether one model achieved a larger point estimate of AUC than another. It is whether the model could isolate a much narrower high-alert subset without erasing too many true positives.

The Brier scores reported for the six models provide another key empirical distinction. Logistic regression, gradient boosting, target-encoded gradient boosting, bagged gradient boosting, and the ensemble all reported Brier score 0.059, whereas naive Bayes reported 0.415. The numerical consequence is large. With prevalence $\pi = 49670 \approx 0.073$ in the held-out sample and 0.076 in the full cohort, a constant-probability forecaster that always predicts the prevalence has Brier error close to $\pi(1 - \pi)$, which is approximately 0.07. A Brier score of 0.059 is therefore better than a prevalence-only predictor. A score of 0.415, by contrast, indicates severe probability mis-scaling [7]. This observation will matter later when the paper turns from interval width to threshold behavior, because it shows that ranking and probability calibration do not move together in this experiment.

The data description also shapes the interpretation of model uncertainty. Low missingness for post-operative white blood cell count, creatinine, and hemoglobin implies that the acute postoperative physiologic variables were measured for most admissions. Higher missingness for preoperative albumin and body mass index implies that some of the slower baseline reserve features were selectively observed. In addition, the reported skewness of white blood cell count, lactate, and creatinine suggests a small subgroup with markedly elevated values concentrated among leak cases, while albumin exhibited left skew with low values overrepresented among positives. These details do not merely describe the data; they help explain why thresholded model behavior is likely to be locally jagged. A small number of admissions sitting in the tail of one or more skewed variables can move a tree-based split or influence the upper tail of a score distribution disproportionately.

The experiment's design choice to use the validation-set F_1 -maximizing threshold should also be treated as part of the empirical frame rather than as a neutral technicality. The F_1 criterion treats false negatives and false positives symmetrically in a harmonic way. That may be acceptable for a comparative proof-of-concept analysis, but it imposes a particular error trade-off. If the cost of a missed leak is much greater than the cost of an unnecessary workup, the clinically desirable threshold would shift. If the opposite is true because alert fatigue or imaging capacity is binding, the threshold would shift differently. The present paper therefore treats the reported threshold not as the true or optimal operating point, but as one observed compromise around which local instability and decision switching can be examined.

A final point concerns what this reconstruction can and cannot do. Since patient-level predicted probabilities are unavailable, the analysis cannot redraw complete ROC or precision-recall curves, cannot compute exact nonparametric confidence intervals for AUC, and cannot locate the full threshold trajectory of each model. It can, however, perform a rigorous finite-sample reading of what the published metrics already imply. That is enough to address the main questions of the paper: how much statistical resolution the test set contains, how fragile the thresholded distinctions are, and how model preference

changes when finite-sample behavior is viewed through explicit error costs rather than only through leaderboard ordering [8].

3. Sampling Resolution and Interval Width

The most immediate limitation of the held-out experiment is numerical resolution. Only forty-nine admissions in the test set belong to the positive class. That means each positive case carries weight $149 \approx 0.0204$ in recall. A single additional retrieved leak raises recall by just over two percentage points; a single missed leak lowers it by the same amount. This is not a criticism of the dataset. It is an unavoidable property of rare outcomes. But it does mean that thresholded sensitivity estimates are inherently coarse. When one model reports recall 0.41 and another reports 0.49, the observed difference corresponds to only about four admissions. In a leaderboard presentation, that contrast can appear more stable than it actually is.

A simple way to display this fact is through standard binomial approximations. If $\hat{r}_m$ is the observed recall of model m on n_1 positive cases and $\hat{p}_m$ is the observed precision based on k_m predicted-positive admissions, then

$$\begin{aligned} \text{SE}\hat{r}_m &= \sqrt{\frac{\hat{r}_m(1 - \hat{r}_m)}{n_1}} \\ \text{SE}\hat{p}_m &= \sqrt{\frac{\hat{p}_m(1 - \hat{p}_m)}{k_m}}. \end{aligned} \quad (3.1)$$

These approximations are imperfect because the thresholds were selected on a validation set and because the k_m values are reconstructed rather than exact, but they are more than adequate for the present purpose: they show the order of magnitude of sampling noise around the observed operating points.

For recall, the interval widths are substantial. Logistic regression, with $\hat{r} = 0.57$, has $\text{SE}\hat{r} \approx 0.071$, implying a rough 95% interval of width about ± 0.14 . Bagged gradient boosting, with $\hat{r} = 0.41$, has $\text{SE}\hat{r} \approx 0.070$, giving a similar width. The three-way ensemble, with $\hat{r} = 0.49$, also lies near ± 0.14 . These are not narrow intervals. They imply substantial overlap among all models except perhaps the naive Bayes configuration at the high-recall end. The practical interpretation is plain: the experiment supports the statement that recall in this family of models lies in a moderate range, but it does not support a sharply resolved ranking among the better configurations based on sensitivity alone [9].

Precision behaves differently because its denominator is the number of predicted-positive admissions rather than the fixed positive count. For logistic regression, which predicts positive for roughly 107 admissions, the standard error of precision is about 0.042, giving a rough 95% width near ± 0.08 . For bagged gradient boosting, which predicts positive for only about 47 admissions, the standard error is about 0.072, and the interval width approaches ± 0.14 . For the three-way ensemble, which predicts positive for about 59 admissions, the corresponding width is roughly ± 0.13 . Thus the model with the highest observed precision does not necessarily have the most stable precision estimate. In fact, because its positive set is small, each additional false alarm or true positive has more leverage on the observed precision. This is precisely the kind of nuance that a point estimate conceals.

These interval widths matter because much of the narrative around the experiment emphasizes precision. The bagged gradient boosting model's precision of 0.43 looks materially better than logistic regression's 0.26, and from an operational perspective that difference is indeed large. Yet the uncertainty around the bagged model's precision is also notably wider. This does not erase the advantage; the gap is large enough that it remains meaningful. What it does undermine is any temptation to read the exact decimal values too literally. A thresholded estimate of 0.43 based on fewer than fifty predicted-positive admissions is not a fixed property of the underlying distribution. It is a sample-level realization with visible spread.

The same issue extends to the F_1 scores implied by the published precision and recall. Since

$$F_{1,m} = \frac{2\hat{p}_m\hat{r}_m}{\hat{p}_m + \hat{r}_m}, \quad (3.2)$$

the estimated F_1 values are approximately 0.357 for logistic regression, 0.206 for naive Bayes, 0.393 for gradient boosting, 0.382 for the target-encoded variant, 0.420 for bagged gradient boosting, and 0.442 for the ensemble. At first glance, the ensemble appears to separate itself from the bagged model by about 0.022 [10]. But that gap is smaller than the magnitude of F_1 perturbation induced by a few event reassignments near threshold, a point that will be formalized in the next section. The observed F_1 ordering among the stronger models is therefore best understood as suggestive rather than exact.

AUC is often treated as if it were less exposed to such instability because it integrates over all thresholds. That is only partly true. The threshold-free nature of AUC avoids some local volatility, but its sampling error can still be substantial when the positive class is small. Under standard large-sample approximations such as the Hanley-McNeil form, an AUC in the mid-0.7 range with $n_1 = 49$ and $n_0 = 621$ carries a standard error on the order of a few hundredths, often around 0.04. This means that the difference between 0.772 and 0.762, or between 0.762 and 0.754, is well within a range where overlap is expected. Even the gap between 0.772 and 0.738, while more noticeable, is still modest relative to the finite-sample resolution available. The experiment therefore supports a family-level observation that the boosting configurations outperform logistic regression in discrimination, but it is much less decisive about the exact internal ranking among the stronger nonlinear models.

The point can be sharpened by focusing on event counts rather than probabilities. The bagged model retrieves about twenty true leaks, the ensemble about twenty-four, the plain gradient boosting model about twenty-five, and logistic regression about twenty-eight. These differences are real, but they are not large on an absolute scale. By contrast, the same models differ in false positives by much wider margins: around 27 for bagged gradient boosting, 35 for the ensemble, 53 for plain gradient boosting, and 80 for logistic regression. In other words, the held-out set is statistically better at distinguishing false-positive accumulation than it is at distinguishing recall among the stronger configurations. That asymmetry is often hidden by the way model performance is summarized.

A Bayesian reading leads to a similar conclusion. If one places a Beta(1, 1) prior on recall and updates with $\widehat{TP}_m$ successes in n_1 positive trials, the posterior for recall is Beta(1 $\widehat{TP}_m$, 1 $\widehat{FN}_m$). The posterior means closely track the observed recalls, but the credible intervals remain wide because n_1 is small [11]. Precision can be treated analogously with k_m predicted-positive admissions as the trial count. The exact posterior intervals differ from the Wald approximations, especially near the boundaries, but the substantive picture does not change: uncertainty around thresholded metrics is large enough that decimal-level distinctions among the better models should not be interpreted as firm hierarchy.

This finite-sample reading does not weaken the overall scientific value of the experiment. Quite the opposite. It clarifies what the experiment establishes robustly. It establishes that the predictive task is tractable but difficult, that the tree-based family offers useful gains in selectivity relative to additive baselines, and that naive Bayes fails badly in probability calibration. What it does not establish with equal confidence is a finely resolved order among every neighboring model configuration. The test set does not have that much granularity. A proper empirical reading should therefore combine the point estimates with an explicit statement of their statistical resolution. Without that, the table can look more definitive than the data permit.

4. Threshold Volatility Around the F1 Optimum

The previous section treated the published metrics as if the operating threshold were fixed. In reality, the threshold itself is estimated. It was selected on the validation set to maximize F_1 , then carried to the held-out set. This procedure is standard, but in a rare-event problem it introduces a second layer of uncertainty. The threshold maximizing F_1 is sensitive to the exact composition of the validation sample

and to the local ordering of a relatively small set of admissions near the cutpoint. When those scores are close, one or two admissions switching sides can change the observed metric balance perceptibly. The result is a threshold that behaves less like a smooth optimum and more like a jagged local maximizer [12].

The source of this jaggedness can be seen directly from the definition of F_1 . For a thresholded classifier,

$$\begin{aligned} F_1 &= \frac{2TP}{2TP \ FP \ FN} \\ &= \frac{2TP}{TP \ FP \ P}, \end{aligned} \quad (4.1)$$

where $P = TP \ FN$ is the total number of positives in the evaluation set. Since P is fixed, F_1 changes only when the threshold reassigns admissions between the four cells of the confusion matrix. But in a low-prevalence problem, these reallocations happen in discrete chunks, not infinitesimal increments. A single positive case crossing the threshold changes TP and FN simultaneously. A single negative crossing the threshold changes FP and TN . Thus the F_1 surface with respect to threshold is piecewise constant with jumps rather than smooth in the ordinary sense.

A local sensitivity calculation makes this concrete. Treating TP , FP , and FN as continuous for approximation, one has

$$\begin{aligned} \frac{\partial F_1}{\partial TP} &= \frac{2FP \ FN}{2TP \ FP \ FN^2} \\ \frac{\partial F_1}{\partial FP} &= -\frac{2TP}{2TP \ FP \ FN^2} \\ \frac{\partial F_1}{\partial FN} &= -\frac{2TP}{2TP \ FP \ FN^2}. \end{aligned} \quad (4.2)$$

These derivatives are best interpreted as the scale of local perturbation rather than literal infinitesimal changes. For the bagged model, with approximate counts $TP = 20.1$, $FP = 26.6$, and $FN = 28.9$, the denominator $2TP \ FP \ FN$ is around 95.7. The resulting local gain from one additional true positive is on the order of 0.012, while the loss from one additional false positive or one additional false negative is on the order of 0.004. The corresponding numbers for the ensemble are of similar magnitude. This matters because the observed F_1 gap between those two models is about 0.022. A very small number of cases near threshold is therefore enough to alter that ordering.

The same point can be made in fully discrete terms [13]. Suppose a single predicted-negative leak is shifted above threshold while the rest of the confusion matrix remains fixed. Then TP increases by 1, FN decreases by 1, and

$$\Delta F_1 = \frac{2TP \ 1}{2TP \ 1 \ FP \ FN - 1} - \frac{2TP}{2TP \ FP \ FN}. \quad (4.3)$$

If instead a single non-leak case is moved above threshold, then FP increases by 1 and

$$\Delta_- F_1 = \frac{2TP}{2TP \ FP \ 1 \ FN} - \frac{2TP}{2TP \ FP \ FN}. \quad (4.4)$$

Evaluated at the operating points of the stronger models, both of these changes are large enough to shift F_1 by several thousandths or more. Since the differences among the better models are themselves

only a few hundredths, one or two admissions around the cutpoint can materially affect the observed ordering.

This threshold volatility is amplified by the rarity of the positive class. Each positive case is a 2.04% unit in recall. Precision is also coarse because the predicted-positive sets are not large. The bagged model’s positive set contains roughly 47 admissions, so each reassignment in that set moves precision by about $1/47 \approx 0.021$ in the crude one-case sense. The ensemble’s positive set contains roughly 59 admissions, implying a corresponding granularity of around 0.017. These are not negligible increments. They mean that the thresholded precision values can change in visible steps even when the underlying score distributions move very little.

The use of validation-set F_1 optimization compounds this problem in a subtle way. The threshold chosen on the validation split is itself a random function of that sample. If one were to repeat the split many times, the selected threshold would vary, especially for models with score distributions that have a flat F_1 region near the optimum. A flat region is not unusual in imbalanced classification, because several adjacent thresholds may yield very similar harmonic trade-offs. The held-out performance then reflects both the intrinsic model behavior and the particular threshold chosen by one finite validation sample [14]. This is another reason to resist overinterpreting the exact decimal values of precision and recall in the final test table.

There is also a model-specific aspect to this instability. Logistic regression typically produces a smoother score distribution because the log-odds are additive and continuous. Tree-based ensembles often produce more clustered scores, especially when the number of terminal regions is moderate. Clustered scores can create threshold plateaus in which many admissions share similar or identical estimated risk. When such a plateau straddles the validation-set optimum, the selected threshold can be especially sensitive to tie-breaking or to which cases happen to sit just above or below the cutpoint. Thus the very model families that improve selectivity through nonlinear partitioning can also exhibit more pronounced threshold discreteness.

This observation should not be misread as a defect specific to trees. It is a structural feature of threshold selection under class imbalance. In fact, the sharper upper-tail separation produced by the better nonlinear models may be clinically beneficial despite threshold jaggedness. The difficulty is interpretive rather than methodological. A thresholded model output should be understood as one draw from a local optimization problem, not as a uniquely determined fixed point. Reporting only a single precision and recall pair at that threshold therefore gives an incomplete account of the experiment. A more informative report would show a narrow threshold neighborhood or, at minimum, the sensitivity of the operating point to small changes in the cutpoint.

One can see the implications by comparing models that are near one another in the leaderboard. The ensemble and bagged models are separated by only a few true positives and false positives. Their relative ranking depends on whether the analyst values peak precision, overall F_1 , PR-AUC, or some other utility [15]. Because the thresholds are themselves estimated with finite noise, the empirical distinction between them should be expressed as a range of plausible operating behaviors rather than a single deterministic choice. The experiment suggests that both occupy an attractive part of the precision-recall frontier. It does not suggest that one has unambiguously solved the threshold problem.

This threshold-centric reading yields a more realistic view of model translation. Clinical systems do not inherit perfect thresholds from development papers. They require calibration, local retuning, and sometimes different error priorities than F_1 . When the threshold chosen in development already sits on a finite-sample plateau, local implementation choices can legitimately move it. For that reason, the threshold volatility revealed here is not just a statistical curiosity. It is part of the bridge from retrospective evaluation to practical use. A model whose apparent superiority depends on a finely balanced threshold may be less robust under deployment than its point estimate suggests. A model with slightly lower headline performance but a broader stable operating region may prove easier to manage. The held-out table alone cannot settle that question, but it can reveal that the question exists.

5. Decision Switching Under Explicit Error Costs

Table 1: Test Set Composition and Class Distribution

Quantity	Value	Description	Proportion
Total admissions (n)	670	Held-out sample size	1.00
Positive cases (n_1)	49	Leak events	0.073
Negative cases (n_0)	621	Non-leak admissions	0.927
Full cohort size	3331	Original dataset	–
Prevalence (overall)	0.076	Leak rate	–

Table 2: Approximate Confusion Matrix Counts by Model

Model	TP	FP	FN
Logistic Regression	27.9	79.5	21.1
Naive Bayes	34.8	255.1	14.2
Gradient Boosting	25.0	53.1	24.0
Target-Encoded GB	21.1	40.9	27.9
Bagged GB	20.1	26.6	28.9
Ensemble	24.0	34.6	25.0

Table 3: Selected Positive Fraction and Alert Rates

Model	Predicted Positives (k_m)	Fraction (α_m)	Interpretation
Logistic Regression	107	0.160	Moderate alert load
Naive Bayes	290	0.433	Very high alerts
Gradient Boosting	78	0.117	Reduced alerts
Target-Encoded GB	62	0.093	Selective
Bagged GB	47	0.070	Highly selective
Ensemble	59	0.087	Balanced

Table 4: Estimated Standard Errors for Recall and Precision

Model	Recall SE	Precision SE	Interval Width (Approx)
Logistic Regression	0.071	0.042	± 0.08 –0.14
Naive Bayes	0.065	0.030	± 0.06 –0.13
Gradient Boosting	0.071	0.050	± 0.10 –0.14
Target-Encoded GB	0.070	0.060	± 0.12 –0.14
Bagged GB	0.070	0.072	± 0.14
Ensemble	0.071	0.065	± 0.13 –0.14

Most comparative model papers stop after reporting thresholded metrics, implicitly assuming that the reader knows how to weigh them. In practice, the relative harm of false negatives and false positives determines which model is preferable. A missed anastomotic leak can be clinically serious. A false alarm may lead to extra imaging, laboratory repetition, senior review, or prolonged monitoring. The

Table 5: F_1 Scores Across Models

Model	Precision	Recall	F_1 Score
Logistic Regression	0.26	0.57	0.357
Naive Bayes	0.12	0.71	0.206
Gradient Boosting	0.32	0.51	0.393
Target-Encoded GB	0.34	0.43	0.382
Bagged GB	0.43	0.41	0.420
Ensemble	0.41	0.49	0.442

Table 6: Switching Thresholds Under Error Cost Ratios

Model Pair	λ^*	Preference Above λ^*	Preference Below λ^*
Bagged vs Ensemble	~ 2.0	Ensemble	Bagged
Bagged vs Logistic	~ 6.7	Logistic	Bagged
Ensemble vs Logistic	~ 11.5	Logistic	Ensemble
GB vs Bagged	~ 5.4	GB	Bagged
Target-Enc vs Bagged	> 14	Target-Enc	Bagged
Naive Bayes vs Bagged	< 2	Naive Bayes	Bagged

Table 7: Projected Alerts per 1000 Admissions

Model	Total Alerts	True Positives	False Positives
Logistic Regression	163	43	120
Naive Bayes	> 400	54	> 350
Gradient Boosting	117	38	79
Target-Encoded GB	93	32	61
Bagged GB	71	31	40
Ensemble	89	37	52

Table 8: Missed Cases per 1000 Admissions

Model	Recall	Missed Cases	Interpretation
Logistic Regression	0.57	33	Lower misses, high alerts
Naive Bayes	0.71	22	High recall, unusable calibration
Gradient Boosting	0.51	37	Moderate trade-off
Target-Encoded GB	0.43	43	Higher misses
Bagged GB	0.41	45	Most economical alerts
Ensemble	0.49	39	Balanced compromise

appropriate balance depends on local capacity, acceptable burden, and the consequence of delay [16]. These considerations can be made explicit by introducing a one-parameter family of linear losses.

Let $L_m \lambda$ denote the thresholded loss of model m , where a false negative costs $\lambda > 0$ units and a false positive costs one unit. Then

$$L_m \lambda = \lambda \widehat{FN}_m \widehat{FP}_m. \quad (5.1)$$

This family does not pretend to represent the full clinical utility of postoperative decision-making. It deliberately ignores true negatives and true positives, treats all false negatives alike, and treats all false

Table 9: Top Predictive Features and Roles

Feature	Type	Role	Data Quality
WBC Count	Lab	Acute response	Low missingness
ICU Admission	Clinical state	Care intensity	Complete
Lactate	Lab	Perfusion marker	Low missingness
Surgery Type	Categorical	Context	Complete
Albumin	Baseline	Nutritional reserve	Higher missingness
Hemoglobin	Lab	Post-op status	Low missingness

Table 10: Model Calibration via Brier Score

Model	Brier Score	Calibration Quality	Interpretation
Logistic Regression	0.059	Good	Reliable probabilities
Naive Bayes	0.415	Poor	Severe miscalibration
Gradient Boosting	0.059	Good	Well-calibrated
Target-Encoded GB	0.059	Good	Stable
Bagged GB	0.059	Good	Reliable
Ensemble	0.059	Good	Consistent

positives alike. Its value lies in transparency. It converts the observed confusion geometry into a clear statement of when one model overtakes another as the false-negative penalty rises.

For any two models a and b , the switching point occurs at

$$\lambda_{a,b}^* = \frac{\widehat{FP}_b - \widehat{FP}_a}{\widehat{FN}_a - \widehat{FN}_b}, \quad (5.2)$$

provided the denominator is nonzero. If $\lambda > \lambda_{a,b}^*$, then the model with fewer false negatives becomes preferable despite any false-positive penalty. If $\lambda < \lambda_{a,b}^*$, the model with fewer false positives is preferred.

This framing immediately clarifies several comparisons. Consider bagged gradient boosting and the three-way ensemble. The bagged model has about 28.9 false negatives and 26.6 false positives. The ensemble has about 25.0 false negatives and 34.6 false positives. Setting the losses equal gives a switching point near $\lambda^* \approx 2.0$. Thus, if one false negative is valued at more than roughly two false positives, the ensemble becomes preferable. If the clinical environment treats false alarms as relatively expensive and false negatives as less than twice as costly, the bagged model is the better choice. This is a much more informative statement than saying that one model had the best AUC and the other had the best PR-AUC.

Now compare bagged gradient boosting with logistic regression. Logistic regression produces fewer false negatives, about 21.1, but many more false positives, about 79.5 [17]. The switching point is approximately $\lambda^* \approx 6.7$. In other words, logistic regression is preferable only if a missed leak is treated as at least six to seven times as costly as an unnecessary flagged non-leak. If the relative cost ratio is lower than that, bagged gradient boosting remains better despite missing more events. This is a striking result because it shows that logistic regression's higher recall does not automatically make it desirable. Its extra sensitivity is purchased through a very large false-positive burden.

The same reasoning can be applied to the ensemble versus logistic regression. The ensemble has both fewer false positives and slightly more false negatives. The switching point is around $\lambda^* \approx 11.5$. That is a high ratio. It means that unless one is willing to treat a missed leak as more than eleven false alarms, the ensemble remains preferable to logistic regression on this simplified loss scale. For many

postoperative surveillance settings, such a high ratio would be difficult to justify, especially if the action triggered by a positive alert includes costly imaging or resource-intensive review.

The comparison between plain gradient boosting and bagged gradient boosting gives a different picture. Plain gradient boosting has about 24.0 false negatives and 53.1 false positives, whereas the bagged model has about 28.9 false negatives and 26.6 false positives. The switching point is near $\lambda^* \approx 5.4$. Thus plain gradient boosting becomes preferable only once the cost of a missed leak is more than five times the cost of a false alarm. The bagged model is therefore the more economical default if alert burden matters. The price of that economy is lower recall.

Target-encoded gradient boosting falls in an even more selective regime. Its false-negative count is close to that of the bagged model, but its false-positive count is larger. The switching point relative to bagged gradient boosting is above $\lambda \approx 14$, which means that the target-encoded variant is difficult to justify under this loss family unless one is extremely averse to false negatives [18]. Since its false-negative reduction relative to bagging is itself very small, its higher false-positive burden overwhelms the comparison under a wide range of plausible cost ratios.

Naive Bayes is dominated in a practical sense by nearly every other configuration once calibration is taken seriously. Even before invoking the catastrophic Brier score, its false-positive count is so large that its higher recall becomes attractive only under very high false-negative penalties. Compared with bagged gradient boosting, the switching point is below two, but that arithmetic ignores the fact that naive Bayes probability estimates are poorly scaled. If the model's risk scores cannot be interpreted numerically, then any threshold-based deployment becomes unstable, and the simple false-negative/false-positive trade-off ceases to capture the full extent of its weakness.

The broader lesson is that pairwise model preference is not a scalar property. It is a function of error cost. The held-out experiment actually contains enough information to draw the switching contours among the better-performing models. Doing so reveals that the bagged and ensemble configurations occupy adjacent but distinct decision regimes. The ensemble is attractive when missed events are moderately more costly than false alarms. The bagged model is attractive when alert economy is more strongly prioritized. Logistic regression only becomes competitive under a very high penalty for misses. This is a more operationally honest interpretation than claiming a single universally best classifier.

One might object that the linear loss family is too crude because it treats all false negatives and false positives as homogeneous. That objection is valid, but it misses the purpose of the exercise. The aim is not to model the full clinical utility function. The aim is to expose the decision assumptions hidden inside performance comparison [19]. The simplified loss family succeeds in doing exactly that. It shows that even in a proof-of-concept setting, model comparison is already a decision problem in disguise. Once made explicit, the structure of that decision becomes easier to reason about.

The same framework also underscores why interval uncertainty matters. The switching points above are computed from approximate confusion counts. Because those counts inherit sampling variability and threshold instability, the exact numerical boundaries are themselves uncertain. But the qualitative ordering is robust. The bagged model occupies the low-burden end. The ensemble shifts toward fewer misses at modest additional burden. Logistic regression sits much farther toward the high-alert side. The loss framework therefore complements the interval calculations of the previous sections. The latter show that some neighboring models are not sharply separated by the data. The former shows that, despite that overlap, the broad geometry of decision preference is still identifiable.

This cost-ratio reading has a practical implication for how empirical papers should be written. Reporting AUC, PR-AUC, precision, and recall is not enough when the models occupy different regions of an error-cost frontier. At minimum, a paper should either declare the assumed relative error costs or show how model preference changes as those costs vary. Without that information, a clinician may mistake a leaderboard for a recommendation [20]. The present experiment is a case in point. Its scientific value is not exhausted by saying that bagged gradient boosting had the highest AUC and the ensemble had the highest PR-AUC. The deeper empirical content is that one model is an alert-economical option and the other is a miss-reducing option, and the preferred choice depends on how the underlying clinical service values those errors.

6. Service-Scale Consequences of Finite-Sample Uncertainty

A thresholded model is usually evaluated on a finite held-out set and then imagined as part of a larger clinical workflow. That transition from test set to service volume is rarely formalized, yet it is exactly where finite-sample uncertainty becomes operationally visible. A recall estimate of 0.49 or a precision estimate of 0.41 is not experienced in practice as an abstract proportion. It becomes numbers of alerts, workups, missed leaks, and false alarms per month or per year. If the held-out set is small, those projected numbers can vary materially even when the point estimate seems stable.

Suppose a surgical service processes N colorectal resections with an event prevalence π similar to the observed cohort. Let r_m and q_m denote the sensitivity and false-positive rate of model m at its chosen threshold. Then the expected number of alerts is

$$\begin{aligned} EA_m &= N\{\pi r_m + 1 - \pi q_m\} \\ &= N\alpha_m, \end{aligned} \tag{6.1}$$

and the expected numbers of true-positive and false-positive alerts are

$$\begin{aligned} ET_m &= N\pi r_m \\ EF_m &= N(1 - \pi q_m). \end{aligned} \tag{6.2}$$

In a deterministic projection, these expectations reproduce the familiar per-1000 summaries. But a service does not receive expected counts; it receives random counts. If one ignores case-mix drift and assumes independent admissions for approximation, then

$$\begin{aligned} \text{Var}A_m &= N\alpha_m(1 - \alpha_m) \\ \text{Var}T_m &= N\pi r_m(1 - r_m) \\ \text{Var}F_m &= N(1 - \pi q_m)(1 - q_m). \end{aligned} \tag{6.3}$$

These variances already produce appreciable spread at moderate service volumes. The held-out uncertainty in r_m and q_m then adds a second layer on top [21].

The service-scale arithmetic is easiest to see for $N = 1000$. Using the reconstructed rates and the cohort prevalence near 7.6%, the bagged model is expected to generate roughly 71 alerts, of which about 31 are true positives and 40 are false positives. The ensemble is expected to generate about 89 alerts, with approximately 37 true positives and 52 false positives. Logistic regression is expected to generate about 163 alerts, with about 43 true positives and 120 false positives. Naive Bayes would generate well over 400 alerts, rendering it operationally unattractive even before its calibration failure is considered.

These point projections are informative, but they can mislead if presented without uncertainty. If the bagged model's recall is approximately 0.41 with a rough 95% width near ± 0.14 , then the expected number of detected leaks per 1000 procedures lies in a broad band rather than at a single value. Similarly, the false-positive rate derived from precision and recall inherits uncertainty from both. A service that plans capacity around the point estimate alone may understate the plausible range of monthly alerts. In a resource-constrained environment, that matters. The difference between forty and sixty false alerts per 1000 cases is not a trivial planning detail.

One can make the propagation explicit with a delta-method approximation. Let $gr, p = \pi r + 1 - \pi p = 1 - \pi$ denote the false-positive rate implied by recall r , precision p , and prevalence π . Then the asymptotic variance of $\hat{q}_m = g\hat{r}_m, \hat{p}_m$ is approximately

$$\begin{aligned}
\text{Var}\hat{q}_m \approx & \left(\frac{\partial g}{\partial r}\right)^2 \text{Var}\hat{r}_m \\
& \left(\frac{\partial g}{\partial p}\right)^2 \text{Var}\hat{p}_m \\
& 2\frac{\partial g}{\partial r}\frac{\partial g}{\partial p}\text{Cov}\hat{r}_m, \hat{p}_m.
\end{aligned} \tag{6.4}$$

The covariance term is not identifiable from the published table, so a conservative descriptive reading can focus on the first two components or regard the omitted covariance as another source of uncertainty. The practical meaning is straightforward. False-positive rate is not directly reported, but once it is derived from two noisy thresholded statistics it can fluctuate noticeably, especially when precision is based on a small predicted-positive set [22].

Service-level uncertainty also highlights the asymmetry between model families. The better nonlinear configurations differ by relatively few true-positive captures per 1000 admissions but by a nontrivial number of false positives. The gap between bagged gradient boosting and the ensemble is roughly six extra true-positive detections at the expense of about twelve additional false alarms per 1000 admissions. Whether that is acceptable depends on local priorities. By contrast, moving from the ensemble to logistic regression adds only about six more true-positive detections but at the expense of roughly sixty-eight more false alarms. Even if the point projections are treated cautiously, the order of magnitude difference in false-alarm expansion is too large to ignore.

Another subtle consequence of finite-sample uncertainty is that the projected workload may be more uncertain for the more selective models than intuition suggests. A model with a small selected fraction produces fewer alerts on average, but each shift in estimated precision or false-positive rate has proportionally greater effect because the upper tail is thin. The bagged model's appealing alert economy is therefore accompanied by more sensitivity of its exact projected counts to small sampling fluctuations. In other words, a narrow predicted-positive set is operationally attractive, but its empirical characterization requires more care because each admission in that set carries higher leverage.

The same reasoning applies to missed cases. Under point estimates, the bagged model misses about 45 leaks per 1000 admissions, whereas the ensemble misses about 39, and logistic regression misses about 33. These are sobering numbers regardless of model choice. They show that none of the reported configurations can function as a low-miss detector on its own. Yet the gap between these values should not be overread. Given the interval width on recall, the difference between missing 39 and missing 45 leaks per 1000 cases is not known with sharp precision from the held-out sample alone. What the experiment supports more strongly is a family-level conclusion: lowering missed-case burden tends to increase alert burden, and the trade-off steepens quickly as one moves away from the more selective configurations [23].

There is also a prevalence issue hidden inside service-scale translation. The held-out set and the full cohort have prevalences of approximately 7.3% and 7.6%, respectively. A real service may differ. Since precision depends directly on prevalence even when recall and false-positive rate do not, the expected fraction of true alerts would shift under a different event rate. A model that looks manageable in a higher-acuity critical care environment could appear much less efficient in a lower-prevalence setting because the same false-positive rate would dilute the alert pool more severely. This observation does not undermine the present analysis; it strengthens the case for treating the reported metrics as conditional on the evaluation setting rather than as universal properties.

A final operational point concerns confidence in escalation pathways. If the practical use of the model is to trigger enhanced surveillance rather than a hard diagnostic decision, then uncertainty around alert counts may be tolerable. If instead the model is intended to drive resource-intensive interventions, such uncertainty becomes more consequential. This is why service-scale propagation belongs in empirical

papers even before implementation is considered. The question is not only whether a model detects signal. It is whether the amount of uncertainty around the detected signal is compatible with the kind of action being contemplated. In the present experiment, the better models likely support a surveillance interpretation more naturally than an invasive decision interpretation, because their expected false-alert volume remains substantial and their miss rates are far from negligible.

7. Measurement Architecture and Feature Dominance

Thresholded uncertainty is only part of the story. The observed model behavior is also shaped by what the data look like. The reported importance ranking and missingness description reveal an observation architecture that helps explain why certain model classes gained selectivity and why one model class failed so badly in calibration. The useful question here is not merely which variables were important, but why their statistical form creates the kind of decision surfaces seen in the experiment [24].

The reported leading features were postoperative white blood cell count, intensive care unit admission status, postoperative lactate, surgery type, and preoperative albumin, with postoperative hemoglobin and several additional baseline or timing variables also contributing. These features are heterogeneous in kind. White blood cell count and lactate are acute physiologic signals measured close to the time of surgery. ICU admission is a care-location state that reflects postoperative acuity, anticipated monitoring need, or both. Surgery type is a categorical description of operative context. Albumin reflects preoperative reserve and nutritional status. Thus the model is not ranking patients on one homogeneous axis. It is integrating acute response, clinical context, baseline reserve, and management intensity.

A convenient notation is to partition the observed feature vector x_i for admission i into four blocks: baseline reserve b_i , surgical context s_i , acute laboratory response ℓ_i , and measurement pattern m_i . Then one may write

$$\begin{aligned} x_i &= b_i, s_i, \ell_i, m_i \\ \eta_i &= f_b b_i \ f_s s_i \ f_\ell \ell_i \ f_m m_i \\ &\quad f_{s\ell} s_i, \ell_i \ f_{\ell m} \ell_i, m_i, \end{aligned} \tag{7.1}$$

where η_i denotes a generic model score. The crucial point is that the cross-terms are likely important. A postoperative lactate elevation means something different in one surgical context than in another. Missing preoperative albumin means something different for an emergent admission than for an elective admission. ICU admission interacts naturally with the laboratory block because it is both a clinical state and a proxy for intensified observation.

This architecture helps explain why naive Bayes failed so sharply in Brier score. Naive Bayes treats predictors as conditionally independent given the class label [25]. Here, that assumption is implausible. White blood cell count and lactate are both manifestations of early physiologic disturbance. ICU admission is likely correlated with deranged laboratory values. Albumin and body mass index are related through nutritional reserve. When such dependencies are ignored, the model can double-count evidence. It may still rank admissions moderately well because the same correlated variables do indicate danger, but its probability scale becomes distorted. That is exactly the pattern reported in the experiment: recall remained high, while the Brier score deteriorated to a level that makes the predicted probabilities almost uninterpretable.

Logistic regression avoids that specific failure because it aggregates information additively on the log-odds scale and does not require conditional independence. Yet additivity imposes its own limitations. If leak risk rises sharply only after white blood cell count exceeds a context-specific threshold, or if the significance of low albumin depends on surgery type, then a purely additive model will blur that structure unless interactions are manually engineered. The reported superiority of the gradient boosting

family over logistic regression is therefore consistent with a data space in which important relationships are thresholded and conditional rather than merely linear.

The skewness pattern reported in the data description reinforces this interpretation. White blood cell count, lactate, and creatinine were right-skewed, with a minority of patients showing markedly elevated values concentrated among leak cases. Albumin was left-skewed, with low values overrepresented among positives. Tree-based learners are naturally suited to such distributions because they can place splits in the tails and isolate relatively small high-risk regions. An additive linear model can capture monotone trends but cannot localize risk as efficiently in a tail-dominated space unless transformed features or interaction terms are added. This is another reason the tree-based models likely improved precision more than recall: they were able to exclude many non-cases from the high-risk region without radically changing the set of positives identified [26].

The measurement-pattern block m_i is equally important. Missingness was not uniform. Routine perioperative laboratory values had low missingness, whereas preoperative and outpatient-linked measurements had higher missingness. Moreover, missingness indicators were included explicitly in the modeling pipeline. This means that the model was allowed to treat absence of information as information. In a critical-care database, such absence can be meaningful. A missing albumin may indicate emergency presentation or incomplete preoperative optimization. A present lactate may indicate enough concern to order it early. The feature space therefore includes both values and patterns of observation. This makes the modeling problem more realistic, but it also increases dependence across blocks because observation intensity and physiologic abnormality are themselves related.

One practical consequence is that the observed model behavior cannot be understood as a pure biological detector. It is detecting a documented postoperative state. That state combines tissue response, surgical context, and how the hospital saw the patient in the first 24 hours. The high rank of ICU admission is particularly instructive. It is not a laboratory marker in the narrow sense. It is a management variable. Its importance suggests that acute postoperative pathway intensity already contains substantial leak-related information [27]. This does not make the model invalid. It clarifies what type of instrument it is: a surveillance-oriented classifier operating on the hospital record as an integrated sensor, not a mechanistic predictor detached from care processes.

The same observation helps interpret the better nonlinear models' selectivity. If the high-risk region in feature space is defined by combinations such as elevated white blood cell count plus ICU admission plus low albumin in a certain procedure subgroup, then a tree-based ensemble can isolate that region compactly. Logistic regression, lacking such local partitioning unless interactions are hand-coded, will spread risk more diffusely across adjacent regions. The result is precisely what the confusion counts suggest: similar families of true positives but very different numbers of false positives. The nonlinear gain is concentrated in exclusion of negatives rather than in wholesale discovery of new positives.

This reading also moderates how the variable ranking should be interpreted. The leading features are not necessarily causal in the intervention sense. They are discriminative within this observation architecture. Postoperative white blood cell count ranks highly because it is both routinely available and strongly associated with early physiologic stress. ICU admission ranks highly because it compresses multiple forms of postoperative concern. Albumin ranks highly because baseline reserve matters and because missingness around it is informative. A high importance score in this setting is best understood as a marker of statistical leverage within a mixed physiologic-observational system.

Seen this way, the experiment provides more than a model comparison. It offers a partial map of where the informative signal resides in the first postoperative day. The signal appears to be strongest when acute physiologic perturbation intersects with care-intensity cues and background reserve [28]. That architecture is consistent with the comparative pattern of the six models. It also helps explain why threshold selection is unstable near the upper tail: a small number of admissions may occupy extreme combinations of these variables, and shifting the threshold around them changes the thresholded metrics in noticeable steps. The measurement architecture and the finite-sample instability therefore reinforce one another rather than describing separate issues.

8. Discussion

The empirical picture that emerges from this analysis is more nuanced than a straightforward model leaderboard. The held-out experiment clearly contains useful signal. It distinguishes a weakly calibrated independence-based model from better-behaved alternatives. It shows that tree-based methods can reduce false-positive accumulation substantially relative to a linear baseline while preserving much of the recoverable signal. It also demonstrates that structured information available within 24 hours after colorectal surgery is not empty with respect to later leak. Those are real findings. But the experiment does not support a highly granular ranking among every neighboring model configuration, nor does it justify reading the published precision and recall values as fixed properties rather than finite-sample realizations.

Three aspects of the evidence deserve particular emphasis. First, the positive class is numerically small enough that thresholded recall estimates are inherently coarse. With forty-nine positive cases in the test set, each event is worth slightly more than two recall points. This means that differences among the stronger models often correspond to only a few admissions. That fact should not lead to dismissal of the results. It should lead to proportionate interpretation. The held-out set is sufficient to establish direction and broad magnitude [29]. It is not sufficient to end all doubt about fine ordering.

Second, the reported F_1 -optimal operating points are much more fragile than a single line in a table suggests. Because F_1 is piecewise constant in threshold with discrete jumps, a small number of cases near the cutpoint can change the observed balance between precision and recall. The magnitude of these local changes is large enough that the F_1 differences among the better nonlinear models could plausibly shift under modest perturbation. Once again, the implication is not that the thresholds are invalid. It is that a threshold chosen on one validation split should be interpreted as one empirical compromise rather than as a uniquely defined optimal setting.

Third, the main experimental separation among models lies in false-positive burden rather than in the number of recovered positives. This is an important result because it changes how “improvement” should be described. The tree-based family does not transform the problem into a high-recall task. Instead, it trims the upper tail, removing many negatives that the linear baseline continues to place above threshold. Bagged gradient boosting pushes this selectivity furthest, while the ensemble gives back some of that economy to recover additional positives. Logistic regression retrieves more positives only by expanding the alert set substantially. That structure is clinically meaningful, and it is visible even after interval uncertainty is acknowledged.

The explicit cost-ratio analysis adds another layer to that interpretation. Different models are preferable under different error priorities. The bagged model is a low-alert option. The ensemble is a moderate-alert option with fewer misses [30]. Logistic regression becomes attractive only when missed leaks are weighted several times more heavily than false alarms. This is not a weakness of the experiment; it is a feature of honest interpretation. A comparative paper should not hide the decision assumptions that turn one metric profile into a preferred model. By writing those assumptions out as switching boundaries, one can make the model comparison more useful to readers who operate in different clinical environments.

The service-scale propagation underscores the same point in operational terms. None of the reported models can be described as a rule-out instrument. Missed cases remain numerous per 1000 admissions even under the better configurations. At the same time, none of the useful models is free of false alarms. The question is therefore not whether the classifier is perfect enough to replace clinical judgment. It is whether a particular precision-recall regime is acceptable for a particular surveillance or escalation task. The finite-sample uncertainty around that regime should be part of the answer, because services plan around volumes, not around abstract proportions.

This paper also suggests a change in how empirical work on rare surgical complications should be reported. The usual combination of AUC, PR-AUC, Brier score, precision, and recall is necessary but incomplete. At minimum, a paper of this kind should also provide interval estimates for thresholded metrics, a threshold neighborhood showing how sensitive the reported operating point is to small changes,

and a simple error-cost analysis showing where model preference switches. None of these additions requires exotic machinery. They are all derivable from the same evaluation framework already used in the experiment. Yet together they yield a much more faithful description of what the data actually support [31].

The observation architecture described by the feature ranking and missingness profile points in the same direction. Since the strongest variables are a mixture of acute physiology, management intensity, and background reserve, the resulting model is inevitably sensitive to how the first postoperative day is documented. That makes local threshold stability even more important, because upper-tail scores may depend on a small number of admissions in which these signals align strongly. It also suggests that future empirical work should pay more attention to score distribution shape and to the stability of selected subsets, not only to global discrimination.

Several limitations of the present paper should be kept in view. The analysis reconstructs confusion counts from rounded precision and recall values, so exact integer counts are unavailable. Complete threshold sweeps and exact confidence intervals for AUC cannot be computed without raw scores. The covariance language used to describe measurement architecture is interpretive rather than estimated directly from patient-level matrices. These constraints matter, but they do not undermine the main conclusions. The finite-sample scale of the test set is known exactly. The thresholded metrics are known to sufficient resolution to show their coarseness. The switching structure under simple error costs follows directly from the reconstructed confusion geometry. Those are the core empirical contributions of the paper.

A final conceptual point is worth stating. In common-outcome settings, the habit of reading model comparisons primarily through point estimates is often tolerable. In rare-event postoperative prediction, that habit is much less defensible. The evaluation sample is too asymmetric, threshold choice is too influential, and service-level consequences are too sensitive to false-positive accumulation [32]. The present experiment is valuable precisely because it exposes those features. Read narrowly, it is a competition among six algorithms. Read more carefully, it is a finite-resolution study of how quickly uncertainty grows when a low-prevalence endpoint is forced into a thresholded decision framework. That broader reading is the one most likely to advance the field.

9. Conclusion

The experiment examined here contains meaningful predictive information, but its most important lessons are not exhausted by the point estimates that headline the model table. With only forty-nine positive cases in the held-out set, recall is coarse, precision for the more selective models is wide, and the F_1 -optimal threshold is a locally unstable object rather than a fixed truth. These conditions do not invalidate the results. They define the resolution at which the results should be interpreted.

Within that resolution, a coherent pattern still appears. Tree-based methods improve selectivity chiefly by reducing false-positive accumulation. Bagged gradient boosting is the most economical of the stronger configurations, whereas the three-way ensemble trades some of that economy for fewer misses. Logistic regression retrieves additional positives only at a much larger alert cost, and naive Bayes is undermined by severe probability mis-scaling. Which model should be preferred depends on how strongly missed leaks are weighted relative to false alarms, and the switching boundaries among the models can be stated explicitly. Rare-outcome surgical prediction should not be reported as though a single leaderboard settled the matter. Interval width, threshold volatility, and cost-ratio dependence are part of the empirical result. Without them, a table of point estimates can appear more decisive than the data allow. With them, the same experiment becomes more informative, because it describes not only what the models achieved, but also how stable, how contingent, and how operationally meaningful those achievements actually are [33].

References

- [1] S. Morarasu, L. O'Brien, C. Clancy, D. Dietrich, C. A. Maurer, M. Frasson, E. García-Granero, and S. T. Martin, "A systematic review and meta-analysis comparing surgical and oncological outcomes of upper rectal, rectosigmoid and sigmoid tumours.," *European journal of surgical oncology : the journal of the European Society of Surgical Oncology and the British Association of Surgical Oncology*, vol. 47, pp. 2421–2428, 5 2021.
- [2] F. N. Wahid and A. E. Ali, "Use of 't-tube' enterostomy in the management of emergency neonatal intestinal problems: a case series," *World journal of pediatric surgery*, vol. 3, pp. e000203–, 11 2020.
- [3] M. I. M. Ilyas, P. Kandagatla, C. A. Reickert, S. P. M. Nalamati, S. Webb, E. J. Szilagyi, and A. J. Stefanou, "Safety of transrectal or transvaginal drainage of pelvic abscesses," *Journal of Coloproctology*, vol. 41, pp. 047–051, 3 2021.
- [4] A. Aryaie, J. L. Singer, M. Fayeizadeh, J. Lash, and J. M. Marks, "Efficacy of endoscopic management of leak after foregut surgery with endoscopic covered self-expanding metal stents (sems)," *Surgical endoscopy*, vol. 31, pp. 612–617, 6 2016.
- [5] B. Sadanandan, "Data-driven risk stratification for anastomotic leak: A proof-of-concept study using electronic health records," *Journal of Data Science, Predictive Analytics, and Big Data Applications*, vol. 9, no. 6, pp. 1–27, 2024.
- [6] J. C. Tham, D. J. Pournaras, B. Alcocer, R. Forbes, A. V. Ariyathenam, M. L. Humphreys, R. G. Berrisford, T. J. Wheatley, D. Chan, G. Sanders, and S. J. Lewis, "Gut hormones profile after an ivor lewis gastro-esophagectomy and its relationship to delayed gastric emptying.," *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus*, vol. 35, 3 2022.
- [7] M. Degiuli, U. Elmore, R. D. Luca, P. D. Nardi, M. Tomatis, A. Biondi, R. Persiani, L. Solaini, G. Rizzo, D. Soriero, D. Cianflocca, M. Milone, G. Turri, D. Rega, P. Delrio, C. Pedrazzani, G. D. D. Palma, F. Borghi, S. Scabini, C. Coco, D. Cavaliere, M. D. Simone, R. Rosati, and R. Reddavid, "Risk factors for anastomotic leakage after anterior resection for rectal cancer (ralar study). a nationwide retrospective study of the italian society of surgical oncology - colorectal cancer network collaborative group.," *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 24, pp. 264–276, 12 2021.
- [8] T. Wang, M. Sadowsky, R. Blakney, P. Coplan, W. Petraiuolo, M. Soberman, J. Tomaszewski, L. Rene, and J. Wood, "Risk of anastomotic leakage with two-row versus three-row manual circular staplers in colorectal anastomosis: a u.s. cohort study.," *International journal of colorectal disease*, vol. 38, pp. 264–264, 11 2023.
- [9] S. Hennig, B. Jansen-Winkel, H. Köhler, L. Knospe, C. Chalopin, M. Maktabi, A. Pfahl, J. Hoffmann, S. Kwast, I. Gockel, and Y. Moulla, "Novel intraoperative imaging of gastric tube perfusion during oncologic esophagectomy-a pilot study comparing hyperspectral imaging (hsi) and fluorescence imaging (fi) with indocyanine green (icg).," *Cancers*, vol. 14, pp. 97–97, 12 2021.
- [10] S. N. Kothari, A. J. Borgert, K. J. Kallies, M. T. Baker, and B. T. Grover, "Long-term (>10-year) outcomes after laparoscopic roux-en-y gastric bypass," *Surgery for obesity and related diseases : official journal of the American Society for Bariatric Surgery*, vol. 13, pp. 972–978, 12 2016.
- [11] S. Knitter, M. M. Maurer, A. Winter, E. M. Dobrindt, P. Seika, P. V. Ritschl, J. Raakow, J. Pratschke, and C. Denecke, "Robotic-assisted ivor lewis esophagectomy is safe and cost equivalent compared to minimally invasive esophagectomy in a tertiary referral center.," *Cancers*, vol. 16, pp. 112–112, 12 2023.
- [12] T. Hossain, A. Jackson, F. Youssef, M. Trotter, F. Yanni, D. Zanutti, J. Catton, N. Welch, R. Vohra, and S. L. Parsons, "Ogc p25 outcomes of laparoscopic vs open oesophagectomy in a single high volume centre," *British Journal of Surgery*, vol. 109, 12 2022.
- [13] S. E. Hwang, M. J. Jung, B. H. Cho, and H. C. Yu, "Clinical feasibility and nutritional effects of early oral feeding after pancreaticoduodenectomy," *Korean journal of hepato-biliary-pancreatic surgery*, vol. 18, pp. 84–89, 8 2014.
- [14] P. Moore, O. Pickering, and E. Cook, "1025 readmission & mortality rates following enhanced recovery after surgery (eras) for colorectal procedures at an island dgh," *British Journal of Surgery*, vol. 108, 5 2021.
- [15] R. Wismayer, "Factors influencing outcome of sigmoid volvulus in northern uganda. a prospective observational study," *East and Central African Journal of Surgery*, vol. 21, pp. 85–97, 1 2017.
- [16] R. R. Gujjuri, S. K. Kamarajah, and S. R. Markar, "Effect of anastomotic leaks on long-term survival after oesophagectomy for oesophageal cancer: systematic review and meta-analysis," *Diseases of the esophagus : official journal of the International Society for Diseases of the Esophagus*, vol. 34, 9 2020.

- [17] O. Aziz, A. Albeyati, M. Derias, N. Varsani, H. Ashrafiyan, T. Athanasiou, S. K. Clark, J. T. Jenkins, and R. H. Kennedy, "Anastomotic leaks can be detected within 5 days following ileorectal anastomosis: a case-controlled study in patients with familial adenomatous polyposis.," *Colorectal disease : the official journal of the Association of Coloproctology of Great Britain and Ireland*, vol. 19, pp. 251–259, 3 2017.
- [18] N. Ersoz, Y. Poyrazoglu, R. Yuksel, B. Uysal, T. Topal, S. K. Tuncer, E. Göçgeldi, and A. Korkmaz, "Poly(adp-ribose) polymerase inhibition improves wound healing in the colonic anastomoses of rats," *European Surgery*, vol. 48, pp. 115–121, 10 2015.
- [19] I. Bartolini, M. Risaliti, M. N. Ringressi, F. Melli, G. Nannini, A. Amedei, P. Muiesan, and A. Taddei, "Role of gut microbiota-immunity axis in patients undergoing surgery for colorectal cancer: Focus on short and long-term outcomes.," *World journal of gastroenterology*, vol. 26, pp. 2498–2513, 5 2020.
- [20] U. Fumagalli, A. Melis, J. Balazova, V. Lascari, E. Morengi, and R. Rosati, "Intra-operative hypotensive episodes may be associated with post-operative esophageal anastomotic leak," *Updates in surgery*, vol. 68, pp. 185–190, 5 2016.
- [21] A. Andreou, M. Biebl, M. Dadras, B. Struecker, I. M. Sauer, P. C. Thuss-Patience, S. Chopra, P. Fikatas, M. Bahra, D. Seehofer, J. Pratschke, and S. C. Schmidt, "Anastomotic leak predicts diminished long-term survival after resection for gastric and esophageal cancer," *Surgery*, vol. 160, pp. 191–203, 4 2016.
- [22] M. Kukar, K. Ben-David, J. S. Peng, K. Attwood, R. M. Thomas, M. Hennon, C. Nwogu, and S. N. Hochwald, "Minimally invasive ivor lewis esophagectomy with linear stapled anastomosis associated with low leak and stricture rates," *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract*, vol. 24, pp. 1729–1735, 7 2019.
- [23] S. Vedantam and J. Roberts, "Endoscopic stents in the management of bariatric complications: Our algorithm and outcomes.," *Obesity surgery*, vol. 30, pp. 1150–1158, 11 2019.
- [24] J. Meyer, H. Joshi, N. C. Buchs, F. Ris, and J. Davies, "Fluorescence angiography likely protects against anastomotic leak in colorectal surgery: a systematic review and meta-analysis of randomised controlled trials.," *Surgical endoscopy*, vol. 36, pp. 7775–7780, 5 2022.
- [25] D. K. Rex, "Smoking increases risk for complications after colorectal surgery," *NEJM Journal Watch*, vol. 2013, 10 2013.
- [26] J. Janež, G. Horvat, A. Jerin, and J. Grosek, "The significance of blood and peritoneal fluid biochemical markers in identifying early anastomotic leak following colorectal resection-findings from a single-center study.," *Medicina (Kaunas, Lithuania)*, vol. 58, pp. 1253–1253, 9 2022.
- [27] S. Scabini, E. Rimini, E. Romairone, R. Scordamaglia, G. Damiani, D. Pertile, and V. Ferrando, "Colon and rectal surgery for cancer without mechanical bowel preparation: One-center randomized prospective trial," *World journal of surgical oncology*, vol. 8, pp. 35–35, 4 2010.
- [28] S. Engelsgerd, S. Kodres-O'Brien, E. Choudhury, B. M. Garijo, J. B. Mason, and K. J. Kowalczyk, "Outcomes and safety of suprapubic vs urethral catheterization following pelvic fascia sparing robotic prostatectomy.," *Urology practice*, vol. 11, pp. 376–384, 12 2023.
- [29] A. R. Bhama, V. Obias, K. B. Welch, J. F. Vandewarker, and R. K. Cleary, "A comparison of laparoscopic and robotic colorectal surgery outcomes using the american college of surgeons national surgical quality improvement program (acsnqip) database," *Surgical endoscopy*, vol. 30, pp. 1576–1584, 7 2015.
- [30] S. Roh, M. D. Iannettoni, J. Keech, E. V. Arshava, A. M. Swatek, M. B. Zimmerman, R. J. Weigel, and K. R. Parekh, "Timing of esophagectomy after neoadjuvant chemoradiation therapy affects the incidence of anastomotic leaks," *The Korean journal of thoracic and cardiovascular surgery*, vol. 52, pp. 1–8, 2 2019.
- [31] S. Ascanelli, G. Navarra, G. Tonini, C. V. Feo, A. Zerbinati, E. Pozza, and P. Carcoforo, "Early and late outcome after surgery for colorectal cancer: elective versus emergency surgery.," *Tumori*, vol. 89, pp. 36–41, 1 2003.
- [32] J. Muniandy, S. R. N. Palaniappan, S. Muthusamy, M. A. Bujang, T. J. Huei, Y. Y. Goh, K. T. T. Long, S. D. Thevendran, and S. S. Sivalingam, "Retrospective cohort study of non-traumatic jejunum and ileum perforation: A multi-center study," *Turkish Journal of Colorectal Disease*, vol. 33, pp. 1–6, 3 2023.
- [33] V. N. Rohith, S. V. Arya, A. Rani, R. K. Chejara, A. Sharma, J. K. Arora, D. S. Kalwaniya, A. Tolat, P. G, and A. Singh, "Preoperative serum albumin level as a predictor of abdominal wound-related complications after emergency exploratory laparotomy.," *Cureus*, vol. 14, pp. e31980–e31980, 11 2022.