Journal of
Personalized Medicine

MDPI

Article

Machine Learning Ensemble to Predict 30-Day Unplanned Reoperation After Anterior Cervical Discectomy and Fusion: Development and Internal Validation in a National Registry

Rohan A. Phadke [1,*,†], Samer G. Salman [1,†], Akhil Marupudi [2], Sri Tummala [3], Zane G. Salman [4], Kirtan Patel [5], Joshua Ong [6], Alireza Tavakkoli [7], Sainyam Galhotra [8], Ajay Tripuraneni [1] and James Rizkalla [9]

1 School of Medicine, Baylor College of Medicine, Houston, TX 77030, USA
2 McGovern Medical School, The University of Texas Health Science Center, Houston, TX 77030, USA
3 College of Medicine, Texas A&M University, Bryan, TX 77807, USA
4 College of Natural Sciences, The University of Texas at Austin, Austin, TX 78712, USA
5 School of Medicine, Edward Via College of Osteopathic Medicine-Louisiana, Monroe, LA 71203, USA
6 Massachusetts Eye and Ear, Harvard Medical School, Boston, MA 02114, USA
7 Department of Computer Science, University of Nevada, Reno, NV 89557, USA
8 Department of Computer Science, Cornell University, Ithaca, NY 14853, USA
9 Department of Orthopaedic Surgery, Baylor University Medical Center, Dallas, TX 75246, USA
* Correspondence: rohan.phadke@bcm.edu
† These authors contributed equally to this work.

Abstract

Background: Existing models for anterior cervical discectomy and fusion (ACDF) predict composite complications, readmission, or reoperation at two years, not early return to the operating room. We developed and validated a machine learning model for 30-day unplanned reoperation using the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) database. **Methods:** We included 68,091 ACDF cases (2015–2021); 1190 (1.75%) had an unplanned reoperation within 30 days. Six algorithms used 64 baseline features, with all preprocessing, calibration, thresholds, and risk tiers derived from training data only. Models were tested on a held-out 30% (n = 20,428; 357 events), with temporal validation (train 2015–2019, test 2020–2021) and five-fold cross-validation. We report the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), calibration, risk tiers, and SHAP (SHapley Additive exPlanations) importance. **Results:** The stacked ensemble (CatBoost and random forest) had an AUROC of 0.712 (95% confidence interval [CI] 0.684–0.741) and an AUPRC of 0.058 (chance 0.017), similar to logistic regression (AUROC 0.711), and calibration was good. The high-risk tier (11.2% of patients) had an observed reoperation rate of 5.27%, 5.4 times higher than that of the low-risk tier (73.8%; 0.98%), but contained only 33.9% of reoperations, whereas 41.5% occurred in the low-risk tier. Sensitivity at the F1-optimal threshold was 23.5%. Temporal validation gave an AUROC of 0.712 (95% CI 0.682–0.743) and cross-validation a mean of 0.709. The strongest predictors were inpatient status, age, American Society of Anesthesiologists (ASA) class, and comorbidity burden. **Conclusions:** Baseline registry data gave modest, temporally stable discrimination and an 11.2% high-risk tier with a 5.4-fold gradient, but only about a quarter to a third of reoperations were detected, and a simpler model performed equally well. The tool suits counseling and surveillance planning, but not ruling out risk. External validation is required before clinical use.

Check for updates

Academic Editor: Davide Bizzoca

Received: 24 October 2026
Revised: 12 November 2026
Accepted: 23 November 2026
Published: 26 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: ACDF; machine learning; ensemble model; risk stratification; unplanned reoperation; NSQIP; SHAP

1. Introduction

Anterior cervical discectomy and fusion (ACDF) is the typically accepted operation for cervical disc pathology, with an estimated 132,000 procedures performed annually in the United States from 2006 to 2013 [1]. Its use has since declined, decreasing in a national claims database from 2010 to 2022 while cervical disc arthroplasty use has increased [2], and falling by 22.9% per 10,000 Medicare fee-for-service beneficiaries from 2013 to 2023 [3]. ACDF nevertheless remains the most common cervical spine operation, accounting for 61.6% of procedures [2]. The procedure is generally safe, but a meta-analysis of 222 studies reported an overall postoperative complication rate of 16% [4].

Reoperation is a central quality metric for ACDF, and early and late reoperations have different drivers. Later reoperation is driven largely by adjacent segment disease and pseudarthrosis: pooled reoperation after ACDF ranged from 3% to 12% at 3 to 7 years of follow-up depending on the study type [5], and in a cohort of 672 patients, 15% underwent revision surgery at a mean follow-up of 31 months, mostly for adjacent segment disease (47.5%) or pseudarthrosis (45.5%) [6]. Early return to the operating room has other causes, most notably postoperative hematoma, which required reoperation within 30 days in 0.40% (about 1 in 250) of 37,261 ACDF patients in the American College of Surgeons National Surgical Quality Improvement Program (ACS-NSQIP) database [7]. In a single-center case-control study, American Society of Anesthesiologists (ASA) physical status class 3 or higher was the only significant risk factor for reintervention within 72 h (odds ratio 5.5, 95% confidence interval 1.1 to 29.9) [8].

Existing risk tools are poorly suited to this endpoint. The ACS-NSQIP Surgical Risk Calculator had poor discrimination for reoperation (area under the curve [AUC] below 0.70) in 253 patients undergoing anterior lumbar interbody fusion [9] and systematically underestimated complication risk in a cohort of 315 spine surgery patients [10]. A published ACDF machine learning study using a claims database predicted 90-day complications, 90-day readmission, and 2-year reoperation, the last with an AUC of 0.671 [11]. A recent scoping review of spine surgery risk models found that traditional scores reach AUCs of roughly 0.64 to 0.78, that machine learning gains over regression are inconsistent and context-dependent, and that limited external validation and calibration drift constrain clinical translation [12]. We therefore developed and validated a machine learning ensemble to predict 30-day unplanned reoperation after ACDF from baseline data, using evaluation metrics suited to a rare outcome and a validation design that keeps all preprocessing, calibration, and threshold selection within the training data.

2. Methods

2.1. Data Source and Cohort

We conducted a retrospective cohort study using the ACS-NSQIP Public Use File from 2015 to 2021. NSQIP is a multi-institutional registry in which trained reviewers record preoperative characteristics, operative details, and 30-day outcomes. In total, 68,091 unique patients who underwent ACDF were included. The primary outcome was unplanned reoperation within 30 days of surgery, defined as an affirmative answer in any of the three NSQIP reoperation fields. This occurred in 1190 patients (1.75%). The Public Use File contains no hospital identifier, so clustering by institution could not be examined.

2.2. Predictors

Only baseline variables recorded before or at surgery were used, yielding 64 features: 38 clinical variables, 10 engineered features, and 16 missingness indicators. Clinical variables were age, sex, race, ethnicity, body mass index (BMI), ASA class (a 1 to 5 rating of overall health), functional status, dyspnea, comorbidities (diabetes, smoking, hypertension, heart failure, chronic obstructive pulmonary disease [COPD], dialysis, disseminated cancer, chronic steroid use, renal failure, and weight loss exceeding 10%), inpatient or outpatient setting, elective status, surgical specialty (orthopedics or neurosurgery), surgery year, wound class, preoperative sepsis, and 12 laboratory values. Laboratory values were limited to the 1st to 99th percentile range, missing values were replaced with the median, and a missingness indicator was added for variables missing in more than 5% of cases. The percentile limits, the medians, and the list of variables receiving missingness indicators were learned in the training set only and then applied unchanged to the tuning and test data. Elective status, renal failure, and weight loss were not recorded in the 2021 file (9950 cases, 14.6%); for that year, elective status was coded as non-elective and the other two variables were imputed. Engineered features were seven binary laboratory flags (creatinine above 1.5 mg/dL, albumin below 3.5 g/dL, hematocrit below 30%, white blood cell [WBC] count above 11,000 cells/μL, international normalized ratio [INR] above 1.5, blood urea nitrogen [BUN] above 20 mg/dL, and sodium below 135 mmol/L), a comorbidity count, a frailty proxy score (0 to 4; functional dependence, weight loss, low albumin, and age above 70 years), and an interaction term between insulin-dependent diabetes and elevated creatinine.

2.3. Model Development

Patients were randomly divided into a 70% training set and a 30% held-out test set, stratified by outcome (Figure 1). One fifth of the training set was reserved to tune hyperparameters and stop training early, and the models were fit on the remaining four fifths. No model, preprocessing step, calibration function, or threshold used the test set. Six models were compared. CatBoost, a gradient-boosted decision tree method, and random forest, an average of many decision trees, were tuned with 80 and 30 Optuna trials (automated hyperparameter search), respectively, to maximize the area under the precision-recall curve (AUPRC; Supplementary Table S6). A multilayer perceptron (MLP) and TabNet, both neural networks, were trained with class weights to offset the rarity of reoperation. A stacked ensemble combined CatBoost and random forest predictions through logistic regression fit on five-fold out-of-fold predictions in the full training set, and a blend ensemble averaged the two with a weight optimized on the same out-of-fold predictions. The CatBoost and random forest models used in the ensembles were refit on the full training set. A class-weighted logistic regression and ASA class alone served as simple reference models. Predicted probabilities were calibrated by Platt scaling (logistic recalibration) fit on out-of-fold training predictions (for the neural networks, on the tuning set).

2.4. Evaluation and Statistical Analysis

Discrimination was measured by the area under the receiver operating characteristic curve (AUROC), defined as the probability that a randomly chosen patient with reoperation receives a higher predicted risk than one without (0.5 is chance; 1.0 is perfect). Because only 1.7% of patients had the outcome, AUPRC, whose chance level equals the event rate, is more informative. We also report the Brier score (mean squared error of predicted probabilities; lower is better) and the calibration slope and intercept (ideal values 1 and 0). Sensitivity, specificity, and positive predictive value were computed at thresholds chosen on out-of-fold training predictions before the test evaluation: the threshold maximizing the F1 score

(which balances precision and sensitivity) and thresholds reaching 50%, 70%, and 80% sensitivity. Ninety-five percent confidence intervals (CIs) for AUROC and AUPRC came from 1000 bootstrap resamples of the test set, and differences between models were assessed using a paired bootstrap on identical resamples. For risk stratification, the stacked ensemble scores were divided into low-, intermediate-, and high-risk tiers at the 75th and 90th percentiles of out-of-fold training scores. Predictor importance was assessed using SHAP values from the CatBoost model in a random sample of 5000 test cases, where the mean absolute SHAP value reflects a variable's average influence on predicted risk. For the stacked ensemble itself, permutation importance (the fall in AUROC when one variable is randomly shuffled) was measured in a random sample of 10,000 test cases. In a sensitivity analysis, performance was repeated in elective, clean-wound patients (excluding disseminated cancer, contaminated or infected wounds, and preoperative sepsis or septic shock) and in non-elective cases. Baseline characteristics were compared using the Mann–Whitney U test (continuous) and chi-square test (categorical). Models were built in Python (version 3.9.6) using CatBoost (version 1.2.10), scikit-learn (version 1.6.1), PyTorch (version 2.8.0), pytorch-tabnet (version 4.1.0), and Optuna (version 4.8.0), and figures were drawn in R (version 4.5.3) using ggplot2 (version 4.0.3).

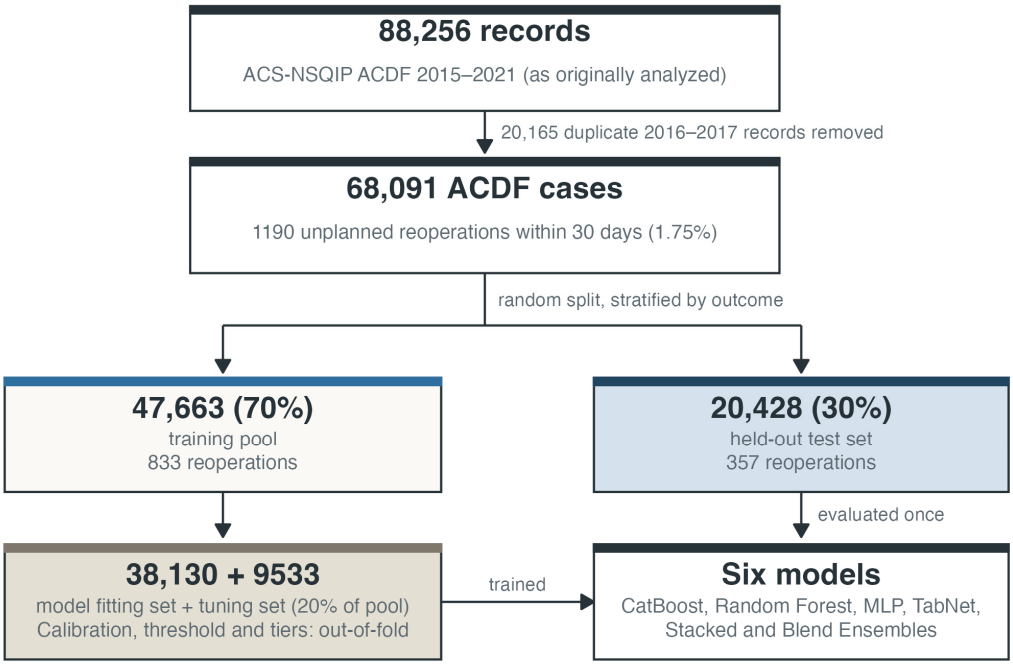

Figure 1. Study flow diagram. ACDF = anterior cervical discectomy and fusion; ACS-NSQIP = American College of Surgeons National Surgical Quality Improvement Program; MLP = multilayer perceptron. Of 88,256 records in the combined file, 20,165 duplicate records were removed, leaving 68,091 cases from 2015–2021 (1190 unplanned reoperations within 30 days, 1.75%), which were randomly split, stratified by outcome, into a 70% training pool (n = 47,663; 833 events) and a 30% held-out test set (n = 20,428; 357 events). Twenty percent of the training pool was set aside to tune hyperparameters and stop training early. Calibration, thresholds, and risk-tier cutoffs came from out-of-fold predictions in the training pool. All performance results come from the test set, which was not used for any fitting, preprocessing, or threshold selection.

2.5. Validation Across Time Periods and Folds

Because the Public Use File lacks hospital identifiers, institution-stratified validation was not possible. Instead, we trained the models on 2015–2019 data and tested them on 2020–2021 data (with and without surgery year as a predictor), used forward chaining (training on all earlier years and testing on 2018, 2019, 2020, and 2021 data separately), and ran

five-fold stratified cross-validation across the full cohort. In every analysis, preprocessing, ensemble stacking, calibration, and tier cutoffs were learned within the training data only. Hyperparameters were those tuned in the main analysis, and stacking used three inner folds.

3. Results

3.1. Cohort

Of 68,091 patients, 1190 (1.75%) had an unplanned reoperation within 30 days (Table 1). Compared with those without reoperation, these patients were older (60.7 vs. 55.8 years), more often male (61.9% vs. 50.2%), more often inpatients (82.8% vs. 60.3%), and less often recorded as elective (66.6% vs. 79.8%; elective status was not recorded for 14.6%, all in 2021) (all $p < 0.0001$). They also had higher ASA class (mean 2.8 vs. 2.5), more insulin-dependent diabetes (10.5% vs. 6.2%) and dialysis (2.2% vs. 0.3%), and lower preoperative albumin (3.9 vs. 4.1 g/dL) (all $p < 0.0001$). Smoking ($p = 0.141$) and surgical specialty ($p = 0.754$) did not differ. The test set contained 20,428 patients and 357 reoperations (Figure 1).

Table 1. Baseline patient characteristics by 30-day unplanned reoperation status.

Variable	Overall (N = 68,091)	No Reoperation (N = 66,901)	Reoperation (N = 1190)	p-Value
Age, years	55.9 ± 11.7	55.8 ± 11.7	60.7 ± 11.4	<0.0001
BMI, kg/m^2	30.7 ± 6.7	30.7 ± 6.7	29.9 ± 6.7	<0.0001
Male sex	34,343 (50.4%)	33,606 (50.2%)	737 (61.9%)	<0.0001
Non-White race	15,550 (22.8%)	15,228 (22.8%)	322 (27.1%)	0.0005
Hispanic ethnicity	3988 (5.9%)	3926 (5.9%)	62 (5.2%)	0.3701
Inpatient setting	41,354 (60.7%)	40,369 (60.3%)	985 (82.8%)	<0.0001
Elective procedure	54,168 (79.6%)	53,376 (79.8%)	792 (66.6%)	<0.0001
Orthopedic specialty	18,524 (27.2%)	18,195 (27.2%)	329 (27.6%)	0.7543
Smoker	16,923 (24.9%)	16,605 (24.8%)	318 (26.7%)	0.1412
Hypertension	32,707 (48.0%)	31,999 (47.8%)	708 (59.5%)	<0.0001
Diabetes (any)	12,219 (17.9%)	11,951 (17.9%)	268 (22.5%)	<0.0001
Insulin-dependent diabetes	4281 (6.3%)	4156 (6.2%)	125 (10.5%)	<0.0001
COPD	3190 (4.7%)	3100 (4.6%)	90 (7.6%)	<0.0001
Heart failure	368 (0.5%)	348 (0.5%)	20 (1.7%)	<0.0001
Dialysis	229 (0.3%)	203 (0.3%)	26 (2.2%)	<0.0001
Disseminated cancer	147 (0.2%)	136 (0.2%)	11 (0.9%)	<0.0001
Chronic steroid use	2606 (3.8%)	2547 (3.8%)	59 (5.0%)	0.0483
Renal failure	53 (0.1%)	44 (0.1%)	9 (0.8%)	<0.0001
Weight loss > 10%	138 (0.2%)	129 (0.2%)	9 (0.8%)	<0.0001
ASA class, mean ± SD	2.5 ± 0.6	2.5 ± 0.6	2.8 ± 0.6	<0.0001
WBC count, 10^3/µL	7.5 ± 2.5	7.5 ± 2.5	8.2 ± 4.2	<0.0001
Hematocrit, %	41.7 ± 4.4	41.8 ± 4.4	40.5 ± 5.6	<0.0001
Creatinine, mg/dL	0.9 ± 0.5	0.9 ± 0.5	1.1 ± 1.0	0.0086
Albumin, g/dL	4.1 ± 0.5	4.1 ± 0.5	3.9 ± 0.7	<0.0001

Continuous variables: mean ± standard deviation (SD); categorical variables: n (%). p-values are from the Mann–Whitney U test (continuous) and chi-square test (categorical). Percentages use all patients as the denominator; elective status, renal failure, and weight loss were not recorded for 9950 patients (14.6%; all 2021 cases), who are counted as not elective or negative. Non-White includes unknown race. Laboratory values use patients with a recorded value. BMI = body mass index; ASA = American Society of Anesthesiologists physical status classification; COPD = chronic obstructive pulmonary disease; WBC = white blood cell.

3.2. Model Performance

The stacked ensemble had the highest AUPRC (0.058, 95% CI 0.045–0.080; chance level 0.017) with an AUROC of 0.712 (95% CI 0.684–0.741) (Table 2, Figure 2), but differences among the top models were small and not statistically distinguishable. The CatBoost model had a nominally higher AUROC (0.713), and the blend ensemble matched the stacked ensemble (AUPRC 0.058, AUROC 0.711); the paired bootstrap difference in AUPRC between the stacked and blend ensembles was 0.000 (95% CI −0.003 to 0.002; Supplementary Table S2).

The stacked ensemble's AUROC exceeded that of the random forest model (0.693) by 0.019 (95% CI 0.008 to 0.031) and TabNet (0.650) by 0.062, but its AUPRC did not differ from that of the random forest or CatBoost models. A class-weighted logistic regression performed equivalently (AUROC 0.711, AUPRC 0.057; paired AUROC difference for the stacked ensemble 0.001, 95% CI −0.014 to 0.016), whereas ASA class alone had an AUROC of 0.618. The stacked ensemble was retained as the primary model, but this designation was made after the test results were seen, and the data do not show that it outperforms simpler models. Its Brier score was 0.0169 (versus 0.0172 for assigning every patient the average risk), and calibration was good (slope 1.04, intercept 0.14; mean predicted risk 1.78% vs. observed 1.75%). At the F1-optimal threshold, sensitivity was 23.5% (95% CI 19.4–28.2), specificity was 93.7%, and the positive predictive value was 6.2%: it flagged 1355 patients (6.6%), 84 of whom had a reoperation, so about one in four reoperations was detected and about 16 patients were flagged for each reoperation found. Thresholds set to reach 50%, 70%, and 80% sensitivity in the training data gave 53.8%, 71.7%, and 80.7% sensitivity in the test set, respectively, while flagging 22.6%, 40.2%, and 52.4% of patients (positive predictive values of 4.2%, 3.1%, and 2.7%; Supplementary Table S1).

Table 2. Performance of six models and two reference models on the held-out test set.

Model	AUROC	AUROC 95% CI	AUPRC	AUPRC 95% CI	Brier	Sensitivity	Specificity	PPV
CatBoost	0.713	[0.685–0.742]	0.057	[0.044–0.077]	0.0169	21.6%	94.0%	6.0%
Random Forest	0.693	[0.663–0.722]	0.054	[0.041–0.077]	0.0169	9.2%	98.2%	8.6%
PyTorch MLP	0.700	[0.673–0.729]	0.057	[0.042–0.080]	0.0169	19.0%	96.2%	8.2%
TabNet	0.650	[0.621–0.681]	0.036	[0.030–0.047]	0.0171	12.3%	96.8%	6.4%
Stacked Ensemble (CB + RF)	0.712	[0.684–0.741]	0.058	[0.045–0.080]	0.0169	23.5%	93.7%	6.2%
Blend Ensemble	0.711	[0.682–0.739]	0.058	[0.045–0.082]	0.0169	20.4%	95.3%	7.2%
Logistic regression (ref.)	0.711	[0.686–0.738]	0.057	[0.042–0.079]	-	-	-	-
ASA class alone (ref.)	0.618	[0.590–0.643]	0.025	[0.022–0.030]	-	-	-	-

Held-out test set (n = 20,428; 357 events). AUROC (area under the receiver operating characteristic curve) and AUPRC (area under the precision-recall curve) are computed on raw model probabilities; Brier score on calibrated probabilities (Platt scaling fit on out-of-fold training predictions; neural networks on the tuning set). A model assigning every patient the average risk has an AUPRC of 0.017 and a Brier score of 0.0172. Sensitivity, specificity, and positive predictive value (PPV) are at the F1-optimal threshold chosen on out-of-fold training predictions (neural networks: tuning set). CIs are 95% bootstrap confidence intervals (1000 resamples). Logistic regression is class-weighted; for ASA class alone, the ASA score is the risk score. The Stacked Ensemble had the nominally highest AUPRC; differences among the top models were not statistically significant. CB = CatBoost; RF = random forest; MLP = multilayer perceptron; - = not applicable.

3.3. Risk Stratification

The stacked ensemble separated patients into three risk tiers (Table 3, Figure 3). The low-risk tier held 73.8% of the test set and had an observed reoperation rate of 0.98%. The intermediate-risk tier (14.9%) had a rate of 2.89%, 2.9 times higher than that of the low-risk tier, and the high-risk tier (11.2%) had a rate of 5.27%, 5.4 times higher than that of the low-risk tier. The tier cutoffs were the 75th and 90th percentiles of out-of-fold training scores, which divided the training data 75%, 15%, and 10% by construction; the test proportions (73.8%, 14.9%, and 11.2%) differed slightly because of sampling variation and small differences between out-of-fold and final-model scores. Mean calibrated predicted risk matched observed risk closely in every tier (1.06% vs. 0.98%, 2.81% vs. 2.89%, and 5.17%

vs. 5.27%). However, the high-risk tier contained only 121 of 357 reoperations (33.9%), whereas 148 (41.5%) occurred in the low-risk tier.

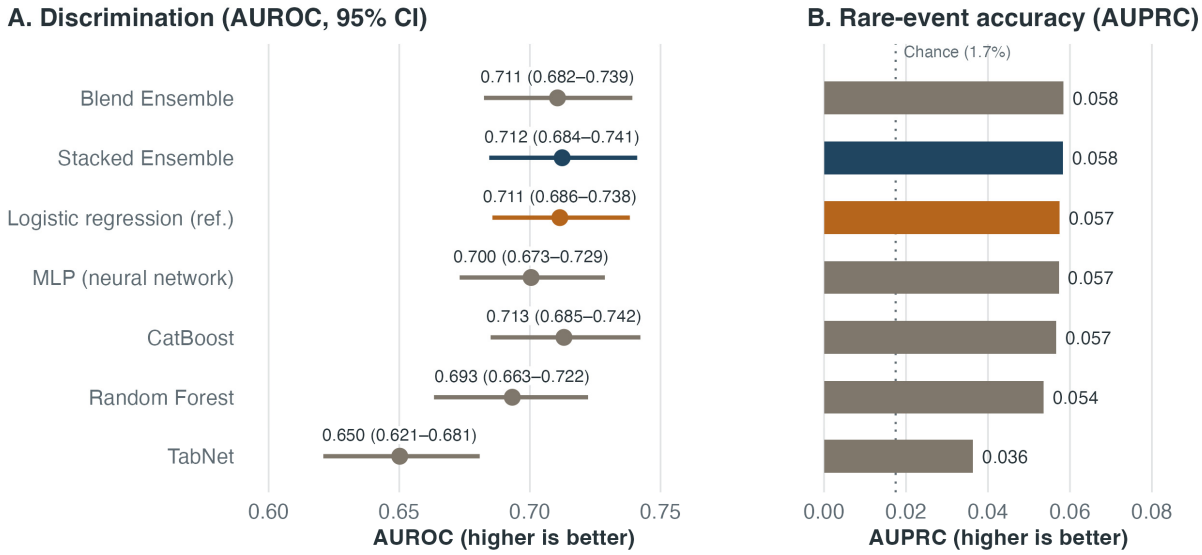

Figure 2. Model performance on the held-out test set (n = 20,428; 357 events). (**A**) Area under the receiver operating characteristic curve (AUROC) with 95% bootstrap confidence intervals (1.0 = perfect separation of patients with and without reoperation; 0.5 = chance). (**B**) Area under the precision-recall curve (AUPRC), which reflects how precisely a model finds the rare reoperations (chance equals the 1.7% event rate, dotted line). Models are ordered by AUPRC; differences among the top models were not statistically significant, and a class-weighted logistic regression performed equivalently (orange). MLP = multilayer perceptron neural network.

Table 3. Risk stratification of the held-out test set by stacked ensemble tier.

Risk Tier	*N* (%)	Reoperations, *n* (% of All)	Observed/Predicted Rate	Relative Risk vs. Low
Low	15,083 (73.8%)	148 (41.5%)	0.98%/1.06%	1.0×
Intermediate	3048 (14.9%)	88 (24.6%)	2.89%/2.81%	2.9×
High	2297 (11.2%)	121 (33.9%)	5.27%/5.17%	5.4×

Held-out test set (n = 20,428). Tier boundaries are the 75th and 90th percentiles of out-of-fold stacked ensemble scores in the training set (cutoffs 0.0224 and 0.0349), applied to the test set. Predicted rates are the mean of Platt-calibrated predicted probabilities. The high-risk tier (11.2% of patients) had an observed reoperation rate of 5.27%, a 5.4-fold difference from the low-risk tier (0.98%), and contained 33.9% of all reoperations; the low-risk tier contained 41.5%.

3.4. Temporal Validation and Cross-Validation

Discrimination was stable over time and across folds (Table 4). Training on 2015–2019 data and testing on 2020–2021 data (18,978 patients; 344 events) gave an AUROC of 0.712 (95% CI 0.682–0.743) and an AUPRC of 0.049 (95% CI 0.041–0.066); excluding surgery year as a predictor gave 0.721 (0.694–0.751) and 0.055. Forward chaining gave AUROC values of 0.695 (2018), 0.719 (2019), 0.732 (2020), and 0.718 (2021). Five-fold cross-validation gave a mean AUROC of 0.709 (SD 0.019; range 0.689–0.733) and a mean AUPRC of 0.056 (SD 0.007; Supplementary Table S3). Across the temporal analyses, the observed rate in the high-risk tier was 3.8 to 5.8 times higher than that in the low-risk tier. Calibration slopes ranged from 0.89 to 1.08, but calibration intercepts shifted (−0.78 to 0.35), indicating that the overall risk level differs between periods and that recalibration would be needed in a new setting.

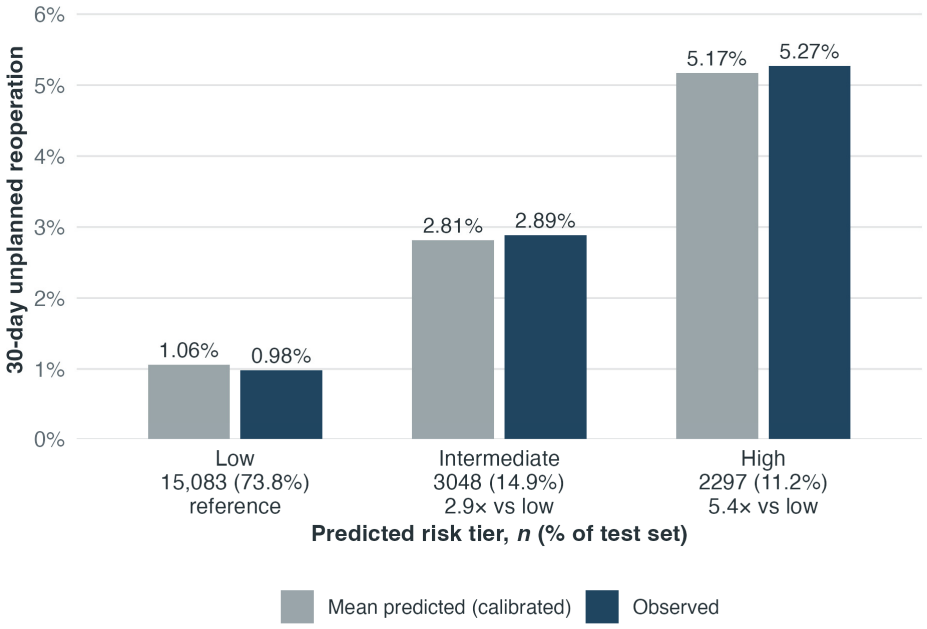

Figure 3. Observed 30-day unplanned reoperation rate and mean calibrated predicted risk by predicted risk tier in the held-out test set. Tiers were defined by the 75th and 90th percentile cutoffs of out-of-fold stacked ensemble scores in the training set. Multipliers show the observed rate relative to the low-risk tier. Patients in the high-risk tier (11.2% of the test set) had a 5.4-fold higher reoperation rate than those in the low-risk tier (73.8%), but the high-risk tier contained only 33.9% of all reoperations.

Table 4. Temporal validation and cross-validation of the stacked ensemble.

Analysis	Test *n*	Events (%)	AUROC (95% CI)	AUPRC (95% CI)	Calibration Slope/Intercept	High vs. Low Tier
Train 2015–2019, test 2020–2021	18,978	344 (1.81%)	0.712 (0.682–0.743)	0.049 (0.041–0.066)	0.94/−0.42	4.7×
Same, without surgery year	18,978	344 (1.81%)	0.721 (0.694–0.751)	0.055 (0.045–0.075)	0.98/−0.24	4.9×
Forward chaining: test 2018	9698	190 (1.96%)	0.695 (0.659–0.728)	0.048 (0.037–0.071)	0.96/0.00	3.8×
Forward chaining: test 2019	10,645	181 (1.70%)	0.719 (0.679–0.756)	0.057 (0.039–0.085)	1.08/0.21	5.0×
Forward chaining: test 2020	9028	169 (1.87%)	0.732 (0.695–0.767)	0.061 (0.047–0.088)	1.08/0.35	5.8×
Forward chaining: test 2021	9950	175 (1.76%)	0.718 (0.675–0.754)	0.054 (0.037–0.081)	0.89/−0.78	4.4×
Stratified 5-fold CV (mean, SD)	13,618	238 (1.75%)	0.709 (SD 0.019; 0.689–0.733)	0.056 (SD 0.007)	0.99/−0.09	5.0×

Each analysis refit all preprocessing, the CatBoost and random forest base models, the stacking meta-learner (three inner folds), calibration, and tier cutoffs within its own training data; hyperparameters were reused from the main analysis. High vs. low = observed reoperation rate in the high-risk tier (top 10% of out-of-fold training scores) divided by that in the low-risk tier (bottom 75%). Calibration intercepts are for Platt-calibrated probabilities (ideal slope 1, intercept 0). Cross-validation row: mean (SD) across five folds, with the range of fold AUROCs. CI = 95% bootstrap confidence interval; AUROC = area under the receiver operating characteristic curve; AUPRC = area under the precision-recall curve.

3.5. Predictors and Sensitivity Analysis

In the CatBoost model, the most influential predictors by mean absolute SHAP value were inpatient status, age, ASA class, comorbidity burden, and male sex, followed by elective

status, the frailty proxy, INR, WBC count, and albumin (Figure 4A). Permutation importance in the stacked ensemble itself gave a similar ordering at the top (inpatient status, ASA class, age, male sex, and the frailty proxy; Figure 4B), with some reordering among lower-ranked variables (Supplementary Table S5). In elective, clean-wound patients (16,155 patients; 230 events, 1.42%), the stacked ensemble AUROC was 0.680 (95% CI 0.642–0.715), lower than in the full test set (0.712; Table 5). In non-elective cases (1209 patients; 65 events, 5.38%), the AUROC was 0.688 (95% CI 0.620–0.748).

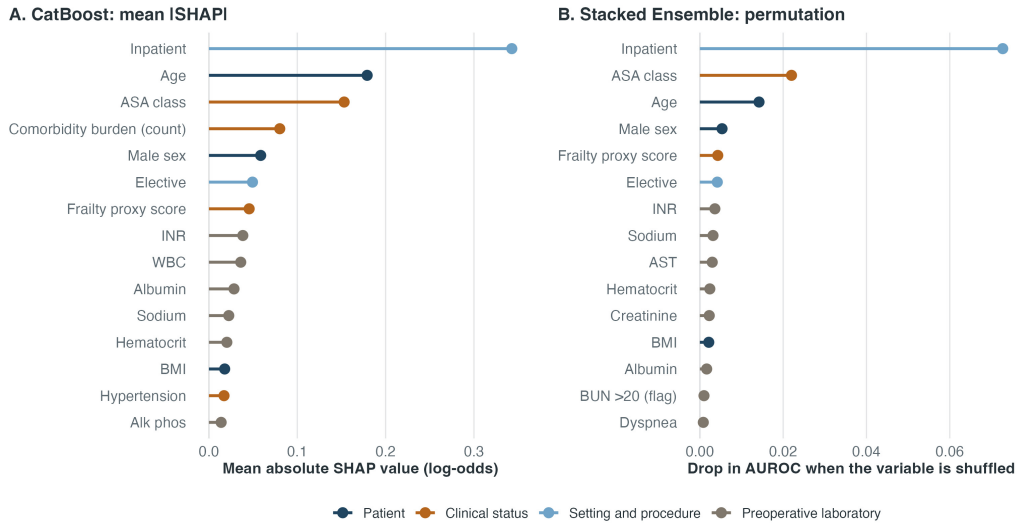

Figure 4. Top 15 predictors in the two models, ranked by importance. (**A**) CatBoost model, ranked by mean absolute SHAP (SHapley Additive exPlanations) value (random sample of 5000 test cases). (**B**) Stacked ensemble, ranked by permutation importance, defined as the decrease in AUROC when a variable is randomly shuffled (random sample of 10,000 test cases). Longer lines indicate a larger influence on predicted risk; the plots show importance, not direction of effect or causation. BMI = body mass index; ASA = American Society of Anesthesiologists; WBC = white blood cell count; BUN = blood urea nitrogen; INR = international normalized ratio; AST = aspartate aminotransferase.

Table 5. Sensitivity analysis by case type (elective, clean-wound and non-elective cases).

Subgroup	N	Events (%)	Stack AUROC (95% CI)	Stack AUPRC	RF AUROC (95% CI)	RF AUPRC
Overall test set	20,428	357 (1.75%)	0.712 (0.687–0.741)	0.058	0.693 (0.663–0.721)	0.054
Elective, clean wound	16,155	230 (1.42%)	0.680 (0.642–0.715)	0.036	0.658 (0.620–0.695)	0.031
Non-elective	1209	65 (5.38%)	0.688 (0.620–0.748)	0.124	0.687 (0.617–0.748)	0.124

The elective, clean-wound cohort (n = 16,155) is elective cases with clean or clean/contaminated wounds, excluding disseminated cancer and preoperative sepsis or septic shock. The non-elective group is cases recorded as not elective. Cases without a recorded elective or non-elective status (3024 in the test set) are in the overall row only. Stack = stacked ensemble; RF = random forest; CI = 95% bootstrap confidence interval (500 resamples); AUROC = area under the receiver operating characteristic curve; AUPRC = area under the precision-recall curve.

4. Discussion

Among 68,091 unique ACDF cases in a national registry, a stacked ensemble of CatBoost and random forest predicted 30-day unplanned reoperation from baseline data with an AUROC of 0.712 and an AUPRC of 0.058, about 3.4 times higher than the chance level of 0.017. Discrimination was modest, was reproduced in temporal validation (AUROC 0.712 for 2020–2021) and cross-validation (mean 0.709), and was matched by a logistic regression using the same variables (0.711), so the ensemble offers no demonstrable advantage over a simpler model. The highest-risk 11.2% of patients had an observed reoperation rate

of 5.27%, 5.4 times higher than that of the low-risk group, which could help focus early postoperative surveillance. However, the tool detected only about a quarter to a third of reoperations (23.5% at the F1-optimal threshold and 33.9% in the high-risk tier), and 41.5% of reoperations occurred in patients classified as low risk.

These results can be set against the existing ACDF prediction literature. Rodrigues et al. applied deep learning to 176,816 ACDF cases in a claims database and reported AUCs of 0.832 for 90-day complications, 0.713 for 90-day readmission, and 0.671 for 2-year reoperation, with logistic regression performing similarly (0.820, 0.712, and 0.671) [11]. Our AUROC of 0.712 for 30-day reoperation is modestly higher than their reoperation estimate, but the data source, follow-up horizon, and outcome definition differ, so the comparison is indirect. A registry-based model in 5060 spine surgery patients reported an AUC of 0.87 for a composite of perioperative complications [13], and artificial neural network and logistic regression models outperformed ASA class alone for cardiac and wound complications, venous thromboembolism, and mortality after posterior lumbar fusion in NSQIP [14]. Those outcomes are composites or different events and are not directly comparable with a single rare outcome. Our discrimination is lower than in the composite-complication models above, as expected for a rare and heterogeneous event such as reoperation. Higher AUROC values have been reported for spine surgery models, such as the 0.87 noted above [13], but these predict composite or more frequent events, and spine prediction models have mostly been validated only internally [12]. Discrimination alone also does not establish clinical applicability, which depends on calibration, the consequences of false-positive and false-negative classifications, external validation, and whether a prediction changes management [12].

The leading predictors (inpatient status, age, ASA class, and the comorbidity burden count) are clinically coherent. ASA class 3 or higher was associated with early reintervention after anterior cervical surgery, in which hematoma was the most frequent cause [8], and with hematoma requiring reoperation, as were multilevel procedures, INR above 1.2, lower BMI, anemia, and male sex [7]. Frailty measured by a modified frailty index independently predicted major complications and reoperation for infection after degenerative spine surgery [15]. In a prospective cervical cohort, however, age and BMI predicted adverse events as well as formal frailty and comorbidity indices, and no index or demographic factor reached an AUC of 0.7 [16]. In our model, age, ASA class, and the engineered comorbidity burden count ranked among the five strongest predictors in the CatBoost SHAP ranking, and the engineered frailty proxy ranked 7th of 64 features (5th by permutation importance in the ensemble), which suggests that composite summaries of routinely recorded comorbidity and function carry some of the signal, although no single index or demographic factor was strongly predictive in prior work [16].

We emphasized AUPRC over AUROC because the outcome was rare. Class imbalance biases the decision boundary of a classifier toward the majority class [17], and in a study of fatal crash prediction with a 0.8% event rate, the area under the precision-recall curve was considered more meaningful than the AUROC and had a random baseline equal to the event rate [18]. Our Brier score of 0.0169 was essentially the same as the 0.0172 of a model that assigns every patient the 1.75% average risk, because for a rare outcome, the Brier score is dominated by the many patients without the event; the calibration slope and intercept and the tier-level comparison are more informative.

SHAP is a game theory method introduced by Lundberg and Lee in 2017 that attributes each prediction to its input features [19] and supports both global and individual-level interpretation [20]. SHAP importance reflects a variable's contribution to the model output, not a causal effect [20]. SHAP values were computed for the CatBoost component, which is one of two base learners in the stacked ensemble. Because the ensemble also contains a

random forest, we additionally computed permutation importance for the stacked ensemble itself, and the leading predictors were consistent between the two methods (Figure 4). The dominant predictors, including age and inpatient status, are not modifiable in the preoperative window, so the tool is better suited to counseling, surveillance, and disposition planning than to risk modification. Retropharyngeal hematoma, one recognized cause of early reoperation, can cause airway compromise and had a median time to formation of 8.5 h (interquartile range, 4 to 24 h) in a series of 10,615 ACDF patients [21]. Intensified monitoring of high-risk patients during the first postoperative day is therefore a plausible use, although it was not tested here. After calibration on out-of-fold predictions, predicted and observed risk agreed closely in all three tiers (5.17% vs. 5.27% in the high-risk tier). Calibration should nevertheless be rechecked in any new population because calibration intercepts shifted between time periods (Table 4), and decile-level estimates are imprecise with only about 36 events per decile (Supplementary Table S4). The ACS-NSQIP calculator discriminated reoperation poorly after anterior lumbar interbody fusion [9], which motivates a procedure-specific model, although we did not compare our model with the calculator. Other NSQIP-based preoperative tools show that admission-day variables alone can support well-calibrated risk tiers, as with the SHORD score for non-home discharge after shoulder arthroplasty (AUROC 0.816; more than 20-fold gradient between the lowest and highest risk groups) [22], although that is a different procedure and outcome.

Several features limit near-term clinical use. First, sensitivity is limited. With a rare outcome, any threshold that keeps the flagged fraction small misses most events (23.5% sensitivity when 6.6% of patients were flagged), whereas capturing about 80% of reoperations required flagging 52.4% of patients with a positive predictive value of 2.7%. The tiers are therefore better described as risk enrichment than as detection, and a low-risk classification should not be reassuring: the low-risk tier still had a 0.98% observed risk and contained 41.5% of all reoperations. Second, the model relies on multidimensional but coarse baseline variables and omits operative details, level, and indication, which probably limits discrimination. Interactions among predictors were available to the tree-based models, yet logistic regression performed equivalently, which suggests little additional signal from interactions in these variables, as also reported for ACDF claims data [11]. Third, calibration and discrimination were stable across years in this registry, but external validation in independent institutions and prospective evaluation of whether risk-guided surveillance changes outcomes are needed before the model is used to counsel patients or allocate monitoring.

4.1. Limitations

Several limitations temper interpretation. First, the 30-day NSQIP window cannot capture the late reoperations that dominate long-term outcomes: pooled reoperation after ACDF ranged from 3% to 12% at 3 to 7 years depending on the study type [5], and symptomatic adjacent segment pathology develops at an estimated 1.6% to 4.2% per year after cervical fusion [23]. The model should not be interpreted as a durability predictor. Second, the Public Use File includes reoperation procedure and diagnosis codes and whether the reoperation was related to the index surgery, but we did not model them, so risk could not be stratified by cause of reoperation, and operative level and indication were not predictors. Chronic steroid use was also a single binary variable that cannot separate corticosteroids from biologics or synthetic disease-modifying drugs; in a 66,162-patient NSQIP ACDF cohort, corticosteroid exposure predicted 30-day infectious complications and readmission but not reoperation [24]. Third, sensitivity was limited: at the F1-optimal threshold, set on out-of-fold training predictions, sensitivity was 23.5%, so roughly three quarters of reoperations were not flagged, and 41.5% of reoperations occurred in the low-risk tier. Consequently, the

tool is better suited to enriching a higher-risk group than to ruling out risk. Fourth, elective status, renal failure, and weight loss were not recorded in the 2021 file (14.6% of cases), so the 2021 portion of the temporal validation lacks this information. Fifth, validation was internal to a single registry. Temporal validation and cross-validation showed stable discrimination, but the Public Use File lacks hospital identifiers, so we could not account for clustering by institution or perform institution-stratified or external validation; differences in care and reporting between hospitals could inflate or deflate performance in new settings, and with 357 test-set events, the confidence intervals are wide. Sixth, the stacked ensemble was designated the primary model after the test results were seen, differences among the top models were not significant, and hyperparameters for the temporal and cross-validation analyses were reused from the main analysis rather than re-tuned. External and ideally prospective validation across institutions and electronic health record-derived populations is required before clinical deployment. A scoping review similarly found that existing spine prediction models are predominantly static and that prospective validation of AI-enabled decision support remains lacking [25]. A scoping review of AI in neurosurgery likewise identified data scarcity and the lack of prospective validation as the main barriers to clinical adoption [26], and even a multidimensional model with excellent internal discrimination, such as a machine learning model of extubation success in 341 children (AUC 0.983 in a single center), was accompanied by a call for prospective multicenter external validation before clinical implementation [27].

4.2. Conclusions

In summary, a stacked ensemble using baseline data produced modest, temporally stable discrimination for 30-day unplanned reoperation after ACDF (AUROC 0.712) and a high-risk tier with a 5.4-fold higher observed rate than the low-risk tier, led by factors such as age and inpatient status that cannot be modified. However, it identified only about a quarter to a third of reoperations and performed no better than logistic regression. Its most appropriate near-term role is risk communication and targeted early surveillance, pending external validation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jpm16120593/s1, Table S1: Operating points of the stacked ensemble on the test set; Table S2: Paired bootstrap differences between models (test set, 1000 resamples); Table S3: Fold-level results of five-fold stratified cross-validation of the stacked ensemble; Table S4: Calibration of the stacked ensemble by decile of predicted risk (test set); Table S5: Predictor importance: CatBoost SHAP and stacked ensemble permutation importance (top 20 by each method); Table S6: Analysis settings.

Author Contributions: Conceptualization, R.A.P. and S.G.S.; methodology, R.A.P. and S.G.S.; data curation, R.A.P. and S.G.S.; formal analysis, R.A.P., S.G.S., J.O., A.T. (Alireza Tavakkoli), S.G. and A.T. (Ajay Tripuraneni); writing–original draft preparation, R.A.P. and S.G.S.; writing–review and editing, R.A.P., S.G.S., A.M., S.T., Z.G.S., K.P. and J.R.; supervision, J.O., A.T. (Alireza Tavakkoli), S.G., A.T. (Ajay Tripuraneni) and J.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study used the deidentified American College of Surgeons National Surgical Quality Improvement Program Public Use File. The work did not constitute human subject research and was exempt from institutional review board review.

Informed Consent Statement: Informed consent was not applicable because the analysis used preexisting deidentified registry data.

Data Availability Statement: The ACS-NSQIP Public Use File is available to participating institutions from the American College of Surgeons. Analysis code is available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Saifi, C.; Fein, A.W.; Cazzulino, A.; Lehman, R.A.; Phillips, F.M.; An, H.S.; Riew, K.D. Trends in resource utilization and rate of cervical disc arthroplasty and anterior cervical discectomy and fusion throughout the United States from 2006 to 2013. *Spine J.* **2018**, *18*, 1022–1029. [CrossRef] [PubMed]
2. Ibrahim, M.T.; Kirven, J.C.; Kavuri, V.; Yu, E.; Singh, V.K. Trends in cervical spine surgery in the United States: A national database analysis. *World Neurosurg.* **2025**, *198*, 123961. [CrossRef] [PubMed]
3. Dinesh, A.; Hirpara, A.; Cheng, C.W.; Furey, C.G.; Rajan, P.V. National trends in anterior cervical discectomy and fusion among Medicare beneficiaries: Declining reimbursement, shrinking workforce, and persistent regional disparities. *Spine* **2026**, *online ahead of print*. [CrossRef] [PubMed]
4. Tavanaei, M.R.; Ansari, M.A.; Hatami, M.A.; Heidari, M.M.J.; Dehghani, M.M.; Hajiloo, P.A.; Khorasanizadeh, M.M. Postoperative complications of anterior cervical discectomy and fusion: A comprehensive systematic review and meta-analysis. *N. Am. Spine Soc. J.* **2025**, *21*, 100596. [CrossRef] [PubMed]
5. Wadhwa, H.; Castro, J.; Wong, H.; Koltsov, J.; Alamin, T. ACDF reoperation and pseudoarthrosis rates are elevated in CDA trials compared to observational studies: Systematic review and meta-analysis. *Eur. Spine J.* **2026**, *online ahead of print*. [CrossRef] [PubMed]
6. van Eck, C.F.; Regan, C.; Donaldson, W.F.; Kang, J.D.; Lee, J.Y. The revision rate and occurrence of adjacent segment disease after anterior cervical discectomy and fusion: A study of 672 consecutive patients. *Spine* **2014**, *39*, 2143–2147. [CrossRef] [PubMed]
7. Bovonratwet, P.; Fu, M.C.; Tyagi, V.; Bohl, D.D.; Ondeck, N.T.; Albert, T.J.; Grauer, J.N. Incidence, risk factors, and clinical implications of postoperative hematoma requiring reoperation following anterior cervical discectomy and fusion. *Spine* **2019**, *44*, 543–549. [CrossRef] [PubMed]
8. Boudissa, M.; Lebecque, J.; Boissière, L.; Gille, O.; Pointillart, V.; Obeid, I.; Vital, J.M. Early reintervention after anterior cervical spine surgery: Epidemiology and risk factors: A case-control study. *Orthop. Traumatol. Surg. Res.* **2016**, *102*, 485–488. [CrossRef] [PubMed]
9. Narain, A.S.; Kitto, A.Z.; Braun, B.; Poorman, M.J.; Curtin, P.; Slavin, J.; Whalen, G.; DiPaola, C.P.; Connolly, P.J.; Stauff, M.P. Does the ACS NSQIP Surgical Risk Calculator accurately predict complications rates after anterior lumbar interbody fusion procedures? *Spine* **2021**, *46*, E655–E662. [CrossRef] [PubMed]
10. Jadresic, M.C.; Baker, J.F. Predicting complications of spine surgery: External validation of three models. *Spine J.* **2022**, *22*, 1801–1810. [CrossRef] [PubMed]
11. Rodrigues, A.J.; Schonfeld, E.; Varshneya, K.; Stienen, M.N.; Staartjes, V.E.; Jin, M.C.; Veeravagu, A. Comparison of deep learning and classical machine learning algorithms to predict postoperative outcomes for anterior cervical discectomy and fusion procedures with state-of-the-art performance. *Spine* **2022**, *47*, 1637–1644. [CrossRef] [PubMed]
12. Salman, S.; Phadke, R.; Kumar, R.; Momin, A.; Tavakkoli, A. Risk prediction in spine surgery: A scoping review of traditional models, artificial intelligence, and the challenge of clinical translation. *Spine Deform.* **2026**, *online ahead of print*. [CrossRef] [PubMed]
13. Campagner, A.; Langella, F.; Bellosta-López, P.; Barile, F.; Cecchinato, R.; Compagnone, D.; Damilano, M.; Lamartina, C.; Redaelli, A.; Vanni, D.; et al. Machine learning-based prediction of perioperative complications in spine surgery: A large-scale model development and validation study. *Eur. Spine J.* **2026**, *online ahead of print*. [CrossRef] [PubMed]
14. Kim, J.S.; Merrill, R.K.; Arvind, V.; Kaji, D.; Pasik, S.D.; Nwachukwu, C.C.; Vargas, L.; Osman, N.S.; Oermann, E.K.; Caridi, J.M.; et al. Examining the ability of artificial neural networks machine learning models to accurately predict complications following posterior lumbar spine fusion. *Spine* **2018**, *43*, 853–860. [CrossRef] [PubMed]
15. Flexman, A.M.; Charest-Morin, R.; Stobart, L.; Street, J.; Ryerson, C.J. Frailty and postoperative outcomes in patients undergoing surgery for degenerative spine disease. *Spine J.* **2016**, *16*, 1315–1323. [CrossRef] [PubMed]
16. Woldu, S.; Solumsmoen, S.; Bech-Azeddine, R. Age and BMI equal modified frailty index, modified Charlson comorbidity index and ASA in predicting adverse events in spinal surgery for cervical degenerative diseases. *Clin. Neurol. Neurosurg.* **2023**, *228*, 107698. [CrossRef] [PubMed]
17. Kim, A.; Jung, I. Optimal selection of resampling methods for imbalanced data with high complexity. *PLoS ONE* **2023**, *18*, e0288540. [CrossRef] [PubMed]
18. Gider, V.; Ekinci, S.; Izci, D.; Gokalp, I.; Rashdan, M.; Salman, M. Comparative evaluation of XGBoost, TabNet, and FT transformer models for fatal crash prediction under extreme class imbalance. *Sci. Rep.* **2026**, *16*, 25809. [CrossRef] [PubMed]

19. Mehdi Chouit, E.; Rachdi, M.; Bellafkih, M.; Raouyane, B. Interpretable machine learning for chronic kidney disease prediction: Insights from SHAP and LIME analyses. *PLoS ONE* **2026**, *21*, e0343205. [CrossRef] [PubMed]
20. Zhu, X.Y.; Li, W.; Pan, X.Y.; Li, T.; Yuan, G.L. Explainable machine learning for long-term cardiovascular disease risk prediction in Chinese middle-aged and older adults: A 9-year longitudinal cohort study with web-based risk calculator. *Sci. Rep.* **2026**, *16*, 14998. [CrossRef] [PubMed]
21. Qu, R.; Jin, J.; Wang, X.; Ma, W.; Zhao, Y.; Pan, S.; Chen, X.; Diao, Y.; Tian, X.; Zhang, F.; et al. The incidence, clinical features, and risk factors for postoperative retropharyngeal hematoma and related dyspnea following anterior cervical discectomy and fusion: A single-center study of 10,615 patients. *Spine J.* **2024**, *24*, 2264–2272. [CrossRef] [PubMed]
22. Phadke, R.A.; Salman, S.G.; Khawaja, S.R.; Fiedler, B.; Rana, A.; Hussain, Z.B.; Garrigues, G.E.; Edwards, T.B. Development and validation of the Shoulder Arthroplasty Risk of Non-Home Discharge (SHORD) score: A risk stratification model for non-home discharge following primary total shoulder arthroplasty. *J. Shoulder Elb. Surg.* **2026**, *online ahead of print*. [CrossRef] [PubMed]
23. Lawrence, B.D.; Hilibrand, A.S.; Brodt, E.D.; Dettori, J.R.; Brodke, D.S. Predicting the risk of adjacent segment pathology in the cervical spine: A systematic review. *Spine* **2012**, *37*, S52–S64. [CrossRef] [PubMed]
24. Phadke, R.A.; Salman, S.G.; Welsh, C.; Lee, N.J. Impact of immunosuppression type on 30-day postoperative outcomes following anterior cervical discectomy and fusion: A retrospective national cohort study of 66,162 patients. *Glob. Spine J.* **2026**, *online ahead of print*. [CrossRef] [PubMed]
25. Salman, S.G.; Phadke, R.; Kumar, R.; Zaman, N.; Tavakkoli, A. Digital twins and multimodal artificial intelligence in spine care: A scoping review of concepts, evidence, and translational barriers. *Spine Deform.* **2026**, *online ahead of print*. [CrossRef] [PubMed]
26. Nasirov, R. Artificial intelligence in neurosurgery: A systematic scoping review of machine learning, deep learning, and large language model applications. *Egypt. J. Neurosurg.* **2026**, *41*, 90. [CrossRef]
27. Sarı, F.; Aliyeva, A. Machine learning-based prediction of 48-hour extubation success in mechanically ventilated children: A single-center retrospective cohort study. *Children* **2026**, *13*, 1166. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.