RESEARCH ARTICLE

Machine Learning Approaches to Forecasting Post-Operative Falls in Elective Total Joint Arthroplasty

Sanjay Chakraborty[1] (s.chakraborty@vcu.edu), Margaret O'Brien[1]

[1]Department of Health Administration, Virginia Commonwealth University, Richmond, VA, USA

Abstract

Patient falls during the early recovery period after total joint arthroplasty can delay rehabilitation and lead to preventable injury. Existing bedside screening tools are practical but rely on a small set of categorical risk factors and may not distinguish risk well when the event rate is low. We evaluated whether supervised machine learning models trained on routinely collected electronic health record data could improve the prediction of documented in-hospital and 30-day post-discharge falls after elective primary hip or knee arthroplasty. The retrospective cohort included 16,408 adult patients treated at a single academic medical center between July 2013 and December 2019. Four models were compared with the Hendrich II Fall Risk Model. They were logistic regression, decision tree, gradient-boosted ensemble, and neural network. Falls were documented for 213 patients (1.30%). To address class imbalance, SMOTE was applied only within the training folds, followed by cross-validated hyperparameter tuning and evaluation on a temporally distinct hold-out set. The gradient-boosted ensemble had the highest test-set area under the receiver operating characteristic curve (AUC = 0.831), followed by the neural network (0.812), logistic regression (0.794), decision tree (0.762), and Hendrich II baseline (0.641). The most influential predictors were comorbidity burden, perioperative medication count, age, body mass index, and benzodiazepine exposure. These results support further external validation of EHR-based fall-risk scores as an adjunct to, rather than a replacement for, clinical assessment in arthroplasty recovery.

Keywords — Predictive analytics; Machine learning; Total joint arthroplasty; Fall prevention; Electronic Health Records

1 Introduction

Total joint arthroplasty (TJA), including total hip arthroplasty (THA) and total knee arthroplasty (TKA), is a common elective operation in the United States. Projections have long indicated rising demand for primary and revision joint replacement through 2030, particularly among patients younger than 65 years [1]. The operation is usually planned, but the recovery period is not risk-free. Falls are a specific concern after arthroplasty because patients resume mobility soon after surgery, often while coping with pain, temporary weakness, altered gait mechanics, and medications that may affect balance or alertness.

The broader burden of falls in older adults is substantial. Bergen, Stevens, and Burns estimated that 29 million falls occurred among adults aged 65 years or older in the United States in 2014, with approximately 7 million injuries and about $31 billion in annual Medicare costs [2]. In arthroplasty populations, inpatient fall rates around 1% have been reported in orthopedic-unit and national-sample studies, while longer post-discharge windows yield higher rates because exposure time is longer [3–5]. Prevention is therefore a balancing problem: restricting mobility may reduce immediate fall exposure, but rehabilitation requires movement; sedating medication may improve comfort, but it can impair balance and attention [6].

Hospitals have traditionally relied on nurse-administered screening tools to stratify fall risk. The Hendrich II Fall Risk Model, validated in a large concurrent case-control study of hospitalized patients, assigns ordinal scores based on confusion, depression, altered elimination, dizziness, gender, administration of antiepileptics or benzodiazepines, and the "Get Up and Go" test [7]. The instrument is simple to use, but it was designed as a general inpatient tool. It does not explicitly model arthroplasty-specific combinations of age, comorbidity burden, medication exposure, procedure type, and secular changes in perioperative care. It also produces a thresholded score rather than an individualized probability.

Electronic health records (EHRs) create an opportunity to estimate risk from information already collected during routine care [8, 9]. Machine learning (ML) methods can model nonlinear relationships and interactions that are difficult to represent in a short bedside score. Related applications include prediction tasks in intensive care and postoperative opioid-prescribing risk after arthroplasty [10, 11]. These models also bring practical and ethical requirements: validation, calibration, interpretability, bias assessment, and clinician acceptance must be addressed before clinical use [12].

This study has two objectives. First, we test whether supervised ML classifiers trained on routinely collected EHR data improve discrimination for post-operative falls among TJA patients compared with the Hendrich II Fall Risk Model. Second, we identify the patient-level variables that contribute most to the best-performing model, enabling clinically meaningful interpretation of the results.

Submitted: September 20, 2023 Accepted: October 7, 2023 Published: November 11, 2023
www.pearlacademicpublishing.com

2 Background and Literature Review

Falls after joint replacement are measured differently across studies. Inpatient events are usually recorded through incident-reporting systems and nursing documentation; post-discharge events depend more heavily on follow-up contact and patient reporting. These differences explain why reported rates vary by setting and follow-up period. An inpatient orthopedic-unit study reported that about 1% of postoperative orthopedic patients fell, and a national study of TJA found that in-hospital fall incidence rose from 0.4% to 1.3% over the study period [3, 4]. Reviews of longer post-discharge windows report higher rates and emphasize older age, prior falls, and functional limitation as recurring risk factors [5].

Several patient-level factors are clinically plausible predictors of falls after arthroplasty: older age, comorbidity burden, obesity, cognitive impairment, prior mobility limitation, and exposure to medications that affect alertness or balance [5, 6]. Medication burden deserves particular attention because postoperative pain regimens may combine opioids, benzodiazepines, muscle relaxants, antiemetics, and other agents. Multimodal analgesia can improve pain management and resource use, but it also increases the need to monitor medication combinations in medically complex patients [13].

The clinical literature contains several fall-risk screening tools developed for the general inpatient population. Among the most widely adopted are the Morse Fall Scale, the Conley Scale, and the Hendrich II Fall Risk Model [7]. These instruments share a common structure: a small set of categorical or ordinal risk factors is scored and summed, and patients exceeding a threshold value are classified as high risk. The Hendrich II model, for example, uses eight items and a cut-off score of five. While these tools are easy to administer and require no computational infrastructure, their discriminatory performance in specialized surgical populations has been called into question. The limited number of predictors, the absence of continuous variable modeling, and the lack of interaction terms constrain the sensitivity and specificity achievable with scorecard approaches.

The application of ML to clinical prediction has expanded rapidly over the past decade, driven by the availability of large-scale EHR data and advances in computational methods [14, 15]. In orthopedic surgery specifically, ML algorithms have been applied to predict length of stay, discharge disposition, readmission, opioid consumption, and patient-reported outcomes following TJA [11, 16, 17]. Across these studies, flexible models such as tree-based ensembles are often competitive with, and sometimes stronger than, traditional regression models when the predictor set includes nonlinearities and interactions.

A recurring methodological challenge in surgical outcome prediction is class imbalance. Adverse events such as falls, surgical site infections, and mortality are, by definition, rare relative to uneventful recoveries. Standard classifiers trained on imbalanced data tend to favor the majority class, resulting in high overall accuracy but poor sensitivity for the minority event of clinical interest [18]. Strategies for addressing this issue include resampling techniques such as the Synthetic Minority Over-sampling Technique (SMOTE), cost-sensitive learning, and threshold adjustment [19]. The choice of evaluation metric is also critical: accuracy alone is misleading when prevalence is low, and area under the receiver operating characteristic curve (AUC), sensitivity, specificity, and the F1-score provide a more informative picture of model performance.

Goldstein and colleagues conducted a systematic review of risk prediction models built with EHR data and identified several common pitfalls, including inadequate handling of missing data, overfitting due to high-dimensional predictor spaces, and failure to validate on temporally distinct cohorts [20]. Our study design incorporates lessons from this review by employing stratified cross-validation, a hold-out temporal validation set, and careful treatment of missing values.

3 Study

3.1 Data

The study cohort was drawn from the EHR system of a single academic medical center in the mid-Atlantic United States. All adults aged 18 years or older who underwent a primary elective THA or TKA between July 1, 2013, and December 31, 2019, were eligible for inclusion. Patients undergoing revision arthroplasty, partial (unicompartmental) knee replacement, or hip hemiarthroplasty were excluded, as were patients with incomplete demographic or surgical records. After exclusions, the final analytic cohort comprised 16,408 patients.

Three linked data tables were extracted from the institutional data warehouse. *Patient demographics and surgical record.* This table contained one row per patient and included gender, age at surgery, body mass index (BMI), year of surgery, primary procedure type (THA or TKA), a binary indicator for whether a post-operative fall occurred, an injury indicator conditional on fall, and fall location (in-hospital vs. post-discharge). Ninety-day complication status, hospital readmission within 90 days, and emergency department (ED) visits within 90 days were retained for descriptive reporting only and were not used as predictors. Each patient was assigned a de-identified sequential key to comply with privacy regulations.

Medication records. A separate table recorded every medication administered or prescribed to each patient during the perioperative period. Each row represented one patient–medication pair. Because most patients received multiple medications, this table contained 182,516 observations. Medications were classified into ther-

apeutic categories, including opioid analgesics, non-steroidal anti-inflammatory drugs, benzodiazepines, anticoagulants, antiemetics, muscle relaxants, corticosteroids, antihypertensives, and supplemental vitamins.

Comorbidity codes. A third table mapped each patient to the set of International Classification of Diseases (ICD) codes recorded in the medical record. There were 3,614 unique comorbidity codes in the dataset. High-frequency comorbidities included osteoarthritis, hypertension, type 2 diabetes mellitus, hyperlipidemia, obesity, gastroesophageal reflux disease, depression, and chronic obstructive pulmonary disease.

The primary outcome was a binary indicator of postoperative fall, defined as any unplanned descent to the floor or a lower level documented in the patient's medical record during the index hospitalization or within 30 days of discharge. Falls were identified through a combination of nursing incident reports and structured follow-up documentation. Of the 16,408 patients in the cohort, 213 (1.30%) experienced at least one documented fall during the observation window.

Table 1 summarizes the demographic and clinical characteristics of the study cohort, stratified by fall status.

Table 1: Demographic and clinical characteristics of the study cohort, stratified by fall status ($N = 16{,}408$).

Characteristic	**No Fall ($n = 16{,}195$)**	**Fall ($n = 213$)**
Demographics		
Female, %	58.0	53.1
Age at surgery, mean (SD)	65.4 (10.2)	72.6 (9.8)
BMI, mean (SD)	31.2 (6.4)	33.7 (7.1)
Procedure		
Total hip arthroplasty, %	42.3	48.8
Total knee arthroplasty, %	57.7	51.2
Comorbidity burden		
Mean number of ICD codes (SD)	5.8 (3.9)	9.4 (4.7)
Hypertension, %	58.2	74.2
Type 2 diabetes, %	22.1	35.7
Depression, %	14.6	27.2
COPD, %	7.3	15.0
Medications (perioperative)		
Mean medication count (SD)	10.4 (4.1)	14.2 (5.3)
Opioid prescribed, %	89.1	95.3
Benzodiazepine prescribed, %	12.8	26.3
Muscle relaxant prescribed, %	18.4	29.1
Post-operative outcomes		
90-day complication, %	6.2	22.5
Readmission within 90 days, %	4.8	18.3
ED visit within 90 days, %	8.1	28.6

3.2 Methodology

Our analytical workflow followed five sequential steps: data preparation, variable analysis and transformation, model selection, model development and training, and model validation and testing.

3.2.1 Step 1: Data Preparation

The three source tables were merged on the de-identified patient key. Because the medication and comorbidity tables contained multiple rows per patient, aggregation was performed prior to the join. For medications, we computed two sets of features per patient: (a) a count of the total number of distinct medications and (b) binary indicators for each of the nine therapeutic categories listed in Section 3. For comorbidities, we computed (a) a count of the total number of unique ICD codes and (b) binary indicators for the 20 most prevalent comorbidity groups, which together covered approximately 87% of all coded conditions. The resulting modeling dataset contained one row per patient and 38 candidate predictor variables. Post-outcome utilization variables were excluded from the predictor matrix to avoid temporal leakage.

Records with missing values in the outcome variable were excluded during cohort construction. Among the predictor variables, BMI was missing for 2.1% of patients. Following guidance from the EHR risk-prediction literature [20], we imputed missing BMI values using the median value within strata defined by age group and gender. No other predictor exhibited a missingness rate above 0.5%.

3.2.2 Step 2: Variable Analysis and Transformation

Continuous variables (age, BMI, medication count, comorbidity count) were examined for distributional properties using histograms and summary statistics. Age and BMI were approximately normally distributed; medication count and comorbidity count were right-skewed. We applied a log transformation to the two count variables to reduce skewness and stabilize variance for the logistic regression and neural network models. The tree-based models do not require distributional assumptions and received the untransformed counts.

Among the categorical predictors, several comorbidity indicators had very low prevalence ($<1\%$ of the cohort). To avoid sparse-cell problems, comorbidity codes with fewer than 100 occurrences were aggregated into an "other comorbidities" indicator. Procedure type was coded as a binary variable (THA vs. TKA), and gender was coded as a binary variable (female vs. male). Surgery year was retained as a categorical variable with seven levels (2013–2019) to capture possible secular trends in surgical practice and fall-prevention protocols.

Variance inflation factors (VIFs) were computed for the continuous predictors to assess multicollinearity. No VIF exceeded 3.2, indicating that collinearity was not a substantial concern.

Because falls occurred in only 1.30% of the cohort, we faced a pronounced class-imbalance problem. In this setting, classifiers trained on the raw distribution tend to predict all patients as non-fallers, achieving apparent accuracy above 98% while failing to identify any actual fall cases [18]. We addressed this issue using SMOTE, which generates synthetic minority-class observations by interpolating between existing minority-class instances in feature space [19]. SMOTE was applied only to the training partition; the validation and test partitions retained the natural class distribution to provide unbiased estimates of real-world performance.

Table 2 provides the complete data dictionary for the analytic dataset.

Table 2: Data dictionary for the analytic dataset.

Variable	**Definition**
Patient Key	De-identified unique sequential identifier
Gender	Female or Male
BMI	Body mass index (kg/m^2)
Age at Surgery	Age in years at date of surgery
Surgery Year	Calendar year of the procedure (2013–2019)
Primary Procedure	THA or TKA
Fall Indicator	Binary: 1 if post-operative fall documented
Injury Indicator	Binary: 1 if fall resulted in documented injury
Fall Location	In-hospital or post-discharge
90-Day Complication	Binary: 1 if any complication within 90 days; descriptive only, not used as a predictor
Readmission	Binary: 1 if readmitted within 90 days; descriptive only, not used as a predictor
ED Visit	Binary: 1 if ED visit within 90 days; descriptive only, not used as a predictor
Comorbidity Count	Number of unique ICD codes per patient
Medication Count	Number of distinct perioperative medications
Comorbidity Indicators	Binary flags for 20 most prevalent conditions
Medication Category Indicators	Binary flags for 9 therapeutic drug categories

3.2.3 Step 3: Model Selection

We selected four supervised classification algorithms that span a range of model complexity and interpretability:

1. **Logistic Regression (LR).** A classical parametric model that estimates log-odds as a linear combination of predictors. LR provides directly interpretable coefficients and serves as a standard benchmark in clinical prediction research.

2. **Decision Tree (DT).** A non-parametric recursive partitioning algorithm that segments the feature space into rectangular regions. Decision trees are transparent and easy to visualize but are prone to overfitting when grown deep.

3. **Gradient-Boosted Ensemble (GBE).** An ensemble method that sequentially fits shallow decision trees to the residuals of the previous iteration, combining many weak learners into a strong classifier. We used the XGBoost implementation with early stopping to control overfitting.

4. **Neural Network (NN).** A feedforward multilayer perceptron with two hidden layers (64 and 32 units, respectively), ReLU activation, and dropout regularization. Neural networks can capture complex nonlinear interactions but are less interpretable than the preceding models.

In addition to the four ML classifiers, we evaluated the Hendrich II Fall Risk Model as a clinical baseline. Hendrich II scores were computed from the available data elements (gender, confusion/disorientation status as proxied by a documented cognitive impairment comorbidity, depression status, altered elimination, dizziness, benzodiazepine prescription, and antiepileptic prescription). Because the "Get Up and Go" test result was not recorded in the EHR, it was coded as zero for all patients, which may slightly underestimate the instrument's performance in prospective use.

Evaluation metrics. Given the low event rate, we adopted AUC as the primary evaluation metric because it is threshold-independent and robust to class imbalance. Secondary metrics included sensitivity (true positive rate), specificity (true negative rate), positive predictive value (PPV), and the F1-score. These were computed at the threshold that maximized the Youden index ($J = \text{sensitivity} + \text{specificity} - 1$).

3.2.4 Step 4: Model Development and Training

The analytic dataset was partitioned using a temporal split. Patients whose surgery occurred between July 2013 and December 2017 ($n = 11{,}237$; 68.5% of cohort) constituted the development set, and patients whose surgery occurred between January 2018 and December 2019 ($n = 5{,}171$; 31.5%) constituted the hold-out test set. Within the development set, five-fold stratified cross-validation was used for hyperparameter tuning and internal validation.

For LR, we used L2 regularization and tuned the regularization strength C over the set $\{0.001, 0.01, 0.1, 1, 10\}$. For DT, we tuned maximum depth (3–15) and minimum samples per leaf (10–100). For GBE, we tuned the number of boosting rounds (100–1,000), maximum tree depth (3–8), learning rate ($\{0.01, 0.05, 0.1\}$), and subsample fraction ($\{0.7, 0.8, 0.9\}$), with early stopping based on validation AUC. For NN, we tuned the dropout rate ($\{0.2, 0.3, 0.4\}$), learning rate ($\{0.0005, 0.001, 0.005\}$), and batch size ($\{64, 128, 256\}$), training for up to 200 epochs with early stopping on validation loss.

Clinical error analysis. In fall prediction, a false negative (Type II error) corresponds to failing to identify a patient who will fall, potentially denying that patient enhanced monitoring and intervention. A false positive (Type I error) corresponds to flagging a patient as high risk when no fall occurs, resulting in unnecessary resource expenditure. We regarded false negatives as more costly than false positives and therefore weighted sensitivity more heavily than specificity in threshold selection.

3.2.5 Step 5: Model Validation and Testing

Final model performance was assessed on the temporally distinct hold-out test set ($n = 5{,}171$). The hold-out set was not used during any stage of model training or hyperparameter tuning. Each model produced a continuous predicted probability, from which a binary classification was derived at the Youden-optimal threshold determined during cross-validation. Confidence intervals for AUC were computed using 2,000 bootstrap replicates.

Table 3: Hyperparameter search grids and selected values for each model.

Model	Parameter	Search Range	Selected
Logistic Regression	Regularization (C)	{0.001, 0.01, 0.1, 1, 10}	0.1
	Penalty type	L2	L2
Decision Tree	Max depth	3–15	7
	Min samples per leaf	10–100	30
Gradient-Boosted Ens.	Boosting rounds	100–1000	450
	Max tree depth	3–8	5
	Learning rate	{0.01, 0.05, 0.1}	0.05
	Subsample fraction	{0.7, 0.8, 0.9}	0.8
Neural Network	Dropout rate	{0.2, 0.3, 0.4}	0.3
	Learning rate	{0.0005, 0.001, 0.005}	0.001
	Batch size	{64, 128, 256}	128

4 Results

Among the 16,408 patients in the analytic cohort, 9,514 (58.0%) were female, and 6,894 (42.0%) were male. The mean age at surgery was 65.6 years (SD = 10.2). The mean BMI was 31.3 (SD = 6.4). Total knee arthroplasty

Table 4: Classification performance of the five models on the hold-out test set ($n = 5{,}171$; 69 falls). Sensitivity, specificity, PPV, and F1 are reported at the Youden-optimal threshold. AUC 95% CI from bootstrap.

Model	AUC	Sensitivity	Specificity	PPV	F1
Hendrich II (baseline)	0.641	0.536	0.710	0.024	0.046
Logistic Regression	0.794	0.710	0.759	0.038	0.072
Decision Tree	0.762	0.667	0.743	0.034	0.065
Neural Network	0.812	0.739	0.766	0.041	0.078
Gradient-Boosted Ens.	**0.831**	**0.768**	**0.774**	**0.044**	**0.084**

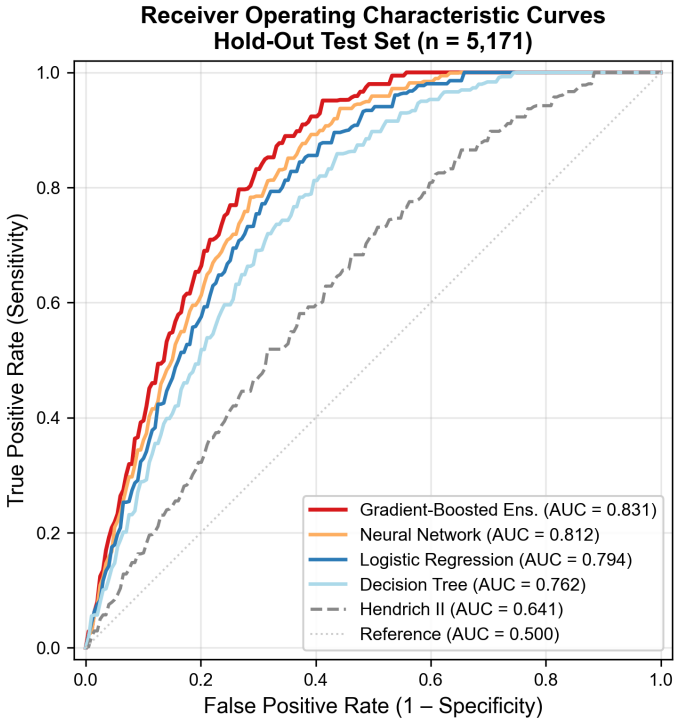

Figure 1: Receiver operating characteristic curves for the five models evaluated on the temporally distinct hold-out test set ($n = 5{,}171$). The gradient-boosted ensemble (GBE) achieves the largest area under the curve.

accounted for 57.6% of procedures and total hip arthroplasty for 42.4%. The overall fall rate was 1.30% ($n = 213$). Patients who experienced a fall were, on average, 7.2 years older, had a 2.5-point higher BMI, carried 3.6 additional comorbidity codes, and were prescribed 3.8 additional perioperative medications compared with patients who did not fall (Table 1).

Table 4 presents the classification performance of all five models on the hold-out test set. The GBE model achieved the highest AUC (0.831; 95% CI: 0.791–0.869), followed by the NN (0.812; 95% CI: 0.770–0.852), LR (0.794; 95% CI: 0.749–0.836), DT (0.762; 95% CI: 0.713–0.808), and the Hendrich II baseline (0.641; 95% CI: 0.588–0.694). All four ML models significantly outperformed the Hendrich II model ($p < 0.001$ for each pairwise DeLong test).

At the Youden-optimal threshold, the GBE model correctly identified 53 of the 69 fall events in the test set (sensitivity = 0.768). Its specificity was 0.774, corresponding to 1,153 false positives and 3,949 true negatives among the 5,102 non-fall patients. The PPV across all models was low (range: 0.024–0.044), reflecting the fundamental arithmetic of rare event prediction: even a highly discriminating classifier will produce many false positives when the base rate is approximately 1%.

Figure 1 presents the receiver operating characteristic (ROC) curves for all five models.

Figure 2 reports the ten most influential predictor variables as determined by the permutation importance method applied to the GBE model. Total comorbidity count was the strongest predictor, followed by the number of distinct perioperative medications, patient age at surgery, BMI, and the presence of a benzodiazepine prescription. These findings are consistent with the clinical intuition that patients with higher medical complexity and greater pharmacological burden face an elevated risk of falls.

Several of the top-ranked variables such as total comorbidity count, perioperative medication count, and BMI,

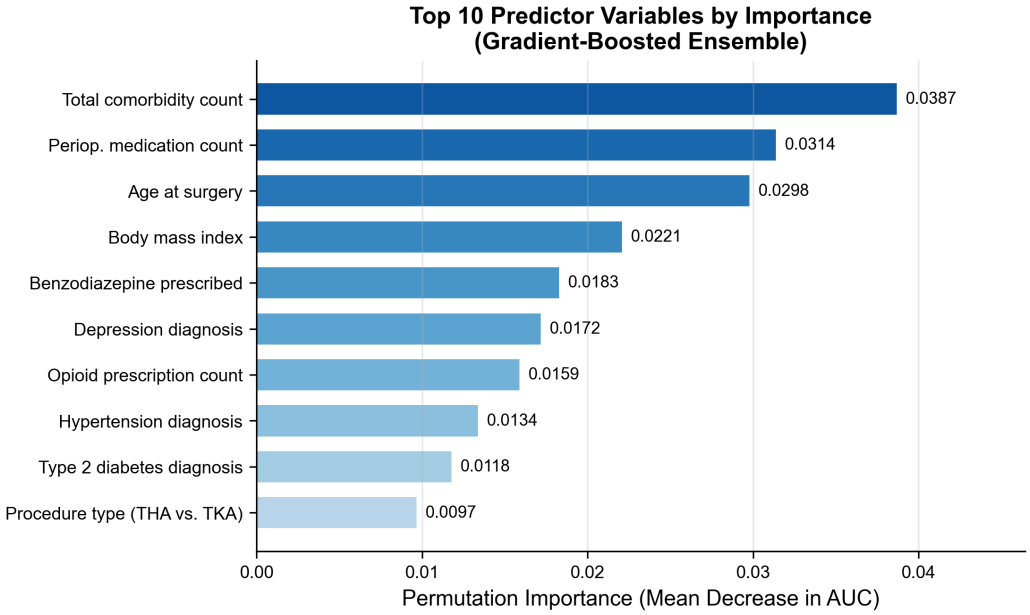

Figure 2: Top ten predictor variables by permutation importance in the gradient-boosted ensemble model. Importance values are the mean decrease in AUC when the variable is randomly permuted.

are continuous measures that are not utilized by the Hendrich II instrument. The ability of the ML models to retain these continuous predictors, rather than reducing them to a few score categories, helps explain the improvement in discrimination over the bedside score.

During five-fold stratified cross-validation on the development set, the GBE model achieved a mean AUC of 0.843 (SD = 0.021). The corresponding figures for LR, DT, and NN were 0.802 (SD = 0.019), 0.774 (SD = 0.028), and 0.826 (SD = 0.024), respectively. The slightly higher cross-validation AUC relative to the hold-out test AUC for most models is consistent with the temporal shift inherent in the train-test split: surgical practice and fall-prevention protocols evolved between the development period (2013–2017) and the test period (2018–2019).

5 Discussion

The central finding of this study is that supervised ML classifiers trained on routinely collected EHR data can substantially outperform the Hendrich II Fall Risk Model at predicting postoperative falls after elective TJA. The GBE model achieved an AUC of 0.831 on a temporally distinct hold-out test set, representing a 19-percentage-point improvement over the Hendrich II baseline (AUC = 0.641). This difference is clinically relevant because fall-prevention resources are limited and usually deployed at the bedside: better risk ranking can help determine which patients need closer monitoring, medication review, or reinforced mobility assistance.

Our results align with the broader trajectory of ML applications in orthopedic surgery. Navarro and colleagues demonstrated that ML models could forecast patient-specific resource utilization after primary TKA with accuracy superior to traditional regression approaches [16]. Ramkumar and colleagues similarly reported that ML-driven risk stratification enabled more effective value-based care pathways for arthroplasty patients at a single institution [17]. Karhade, Schwab, and Bedair showed that gradient-boosted methods outperformed logistic regression in predicting sustained opioid prescriptions after THA [11]. Our findings apply the same general approach to the narrower outcome of post-operative falls.

The variable importance analysis offers actionable clinical insight. Total comorbidity count and perioperative medication count emerged as the two strongest predictors, underscoring the role of medical complexity in fall risk. These variables represent the cumulative physiological burden that a patient carries into surgery and the pharmacological complexity of the recovery period. The prominence of benzodiazepine prescription in the top five predictors corroborates earlier work linking sedative-hypnotic medications to fall risk in older adults [6]. Depression, ranked sixth, should be interpreted cautiously. It may capture several mechanisms at once, including medication exposure, lower activity, worse baseline function, or differences in documentation, rather than a single causal pathway.

The low PPV observed across all models (range: 0.024–0.044) warrants careful interpretation. When the base rate is 1.3%, even a well-performing classifier will produce many more false positives than true positives at any clinically useful sensitivity level. However, the PPV should not be interpreted in isolation. From a clinical perspective, the relevant comparison is between the ML-derived risk score and the current standard of care

(Hendrich II), not between the ML score and a hypothetical perfect classifier. At comparable sensitivity levels, the GBE model generates substantially fewer false positives than the Hendrich II model, thereby enabling the additional monitoring resources to be allocated more efficiently. The interventions usually triggered by a high-risk classification, closer observation, assistive device review, medication review, and patient education, are comparatively low risk, but they still consume staff time. Threshold selection should therefore be tied to local staffing and prevention protocols rather than fixed solely by the Youden index.

The ethical dimensions of deploying ML-based fall prediction in clinical settings deserve consideration. Char, Shah, and Magnus have cautioned that ML systems can encode and perpetuate existing disparities in care delivery, particularly when trained on historical data that reflect inequitable access or biased documentation practices [12]. In our study, the cohort is drawn from a single academic medical center and may not represent the demographic, socioeconomic, or clinical diversity of the broader arthroplasty population. External validation across multiple institutions and careful monitoring for algorithmic bias would be necessary preconditions for clinical deployment.

This study has four main limitations. First, the retrospective, single-institution design limits the generalizability of the findings. The patient population may differ systematically from populations served by community hospitals, rural facilities, or institutions in other geographic regions. External validation on independent datasets is essential before the models can be recommended for widespread clinical use. Second, the outcome definition relied on documented fall events in the medical record. Falls that occurred after discharge but were not reported to the treating institution would not have been captured, leading to possible outcome misclassification. This ascertainment bias would tend to attenuate the observed associations and underestimate the true fall rate. Third, the Hendrich II baseline was computed without the “Get Up and Go” test component, because this assessment was not routinely recorded in the EHR. This omission may have disadvantaged the Hendrich II model relative to its performance in prospective clinical use, where the test is administered at the bedside. Fourth, although SMOTE is a widely used technique for addressing class imbalance [19], it introduces synthetic observations that may not fully represent the distribution of real minority-class cases, particularly in high-dimensional feature spaces. Alternative approaches, such as cost-sensitive learning or ensemble-level resampling, could be explored in future work.

6 Conclusion

In this retrospective single-center cohort, supervised machine-learning classifiers trained on EHR data improved discrimination for documented postoperative falls after elective total joint arthroplasty compared with the Hendrich II Fall Risk Model. Among the four algorithms evaluated, the gradient-boosted ensemble achieved the strongest discriminatory performance (AUC = 0.831), followed by the neural network, logistic regression, and decision tree. Comorbidity burden, polypharmacy, patient age, body mass index, and benzodiazepine use were the most influential predictors, highlighting the interplay between medical complexity and pharmacological exposure in determining fall risk.

The projected growth in joint replacement volume [1] strengthens the case for accurate, scalable risk-prediction tools. Machine learning offers a promising pathway toward individualized perioperative risk stratification that can be integrated into clinical decision support systems within existing EHR infrastructure [8, 14]. Future work should pursue multi-institutional external validation, prospective clinical trials assessing the impact of model-guided interventions on fall rates, and exploration of deep learning architectures that incorporate unstructured clinical text.

References

[1] S. M. Kurtz, E. Lau, K. Ong, K. Zhao, M. Kelly, and K. J. Bozic, “Future young patient demand for primary and revision joint replacement: National projections from 2010 to 2030,” *Clinical Orthopaedics and Related Research*, vol. 467, no. 10, pp. 2606–2612, 2009.

[2] G. Bergen, M. R. Stevens, and E. R. Burns, “Falls and fall injuries among adults aged $\geq$65 years—United States, 2014,” *MMWR. Morbidity and Mortality Weekly Report*, vol. 65, no. 37, pp. 993–998, 2016.

[3] D. B. Ackerman, R. T. Trousdale, P. Bieber, J. Henely, M. W. Pagnano, and D. J. Berry, “Postoperative patient falls on an orthopedic inpatient unit,” *Journal of Arthroplasty*, vol. 25, no. 1, pp. 10–14, 2010.

[4] S. G. Memtsoudis, C. J. Dy, Y. Ma, Y.-L. Chiu, A. G. D. Valle, and M. Mazumdar, “In-hospital patient falls after total joint arthroplasty: Incidence, demographics, and risk factors in the united states,” *Journal of Arthroplasty*, vol. 27, no. 6, pp. 823–828.e1, 2012.

[5] Y. Liu, Y. Yang, H. Liu, W. Wu, X. Wu, and T. Wang, “A systematic review and meta-analysis of fall incidence and risk factors in elderly patients after total joint arthroplasty,” *Medicine*, vol. 99, no. 50, p. e23664, 2020.

[6] M. E. Tinetti and C. Kumar, "The patient who falls: "It's always a trade-off"," *JAMA*, vol. 303, no. 3, pp. 258–266, 2010.

[7] A. L. Hendrich, P. S. Bender, and A. Nyhuis, "Validation of the Hendrich II fall risk model: A large concurrent case/control study of hospitalized patients," *Applied Nursing Research*, vol. 16, no. 1, pp. 9–21, 2003.

[8] D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, and G. Escobar, "Big data in health care: Using analytics to identify and manage high-risk and high-cost patients," *Health Affairs*, vol. 33, no. 7, pp. 1123–1131, 2014.

[9] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *New England Journal of Medicine*, vol. 375, no. 13, pp. 1216–1219, 2016.

[10] D. Shillan, J. A. C. Sterne, A. Champneys, and B. Gibbison, "Use of machine learning to analyse routinely collected intensive care unit data: A systematic review," *Critical Care*, vol. 23, no. 1, p. 284, 2019.

[11] A. V. Karhade, J. H. Schwab, and H. S. Bedair, "Development of machine learning algorithms for prediction of sustained postoperative opioid prescriptions after total hip arthroplasty," *Journal of Arthroplasty*, vol. 34, no. 10, pp. 2272–2277.e1, 2019.

[12] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care—addressing ethical challenges," *New England Journal of Medicine*, vol. 378, no. 11, pp. 981–983, 2018.

[13] S. G. Memtsoudis, J. Poeran, N. Zubizarreta, C. Cozowicz, E. E. Mörwald, E. R. Mariano, and M. Mazumdar, "Association of multimodal pain management strategies with perioperative outcomes and resource utilization: A population-based study," *Anesthesiology*, vol. 128, no. 5, pp. 891–902, 2018.

[14] A. Rajkomar, J. Dean, and I. S. Kohane, "Machine learning in medicine," *New England Journal of Medicine*, vol. 380, no. 14, pp. 1347–1358, 2019.

[15] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: Past, present and future," *Stroke and Vascular Neurology*, vol. 2, no. 4, pp. 230–243, 2017.

[16] S. M. Navarro, E. Y. Wang, H. S. Haeberle, M. A. Mont, V. E. Krebs, B. M. Patterson, and P. N. Ramkumar, "Machine learning and primary total knee arthroplasty: Patient forecasting for a patient-specific payment model," *Journal of Arthroplasty*, vol. 33, no. 12, pp. 3617–3623, 2018.

[17] P. N. Ramkumar, H. S. Haeberle, M. R. Bloomfield, J. L. Schaffer, A. F. Kamath, B. M. Patterson, and V. E. Krebs, "Artificial intelligence and arthroplasty at a single institution: Real-world applications of machine learning to big data, value-based care, mobile health, and remote patient monitoring," *Journal of Arthroplasty*, vol. 34, no. 10, pp. 2204–2209, 2019.

[18] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[20] B. A. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review," *Journal of the American Medical Informatics Association*, vol. 24, no. 1, pp. 198–208, 2017.