

ORIGINAL RESEARCH ARTICLE

Beyond SMOTE: Evaluating large language models and mixture of experts for prediction of surgical site infections

Stephen Russell*

Department of Intelligent Systems and Robotics, University of West Florida, Pensacola, Florida, United States of America

Abstract

Handling severe class imbalance remains one of the most persistent barriers to deploying reliable artificial intelligence (AI) in healthcare. Conventional approaches such as SMOTE and other resampling strategies often inflate training performance but degrade under real-world distribution shift. We evaluated alternative modeling strategies, including classical machine learning, ensemble methods, imbalance-aware mixture of experts (MoE), and a fine-tuned large language model (LLM) (ModernBERT-large), for surgical site infection (SSI) prediction using structured electronic medical records. Across temporally shifted evaluation cohorts, the ModernBERT model consistently outperformed all baselines without synthetic oversampling or target ratio adjustments, achieving a Matthews correlation coefficient of 0.71 versus 0.35 for the best SMOTE-resampled CatBoost model. In contrast, MoE architectures failed to deliver robustness gains, and resampled classical models deteriorated under distributional change. These results highlight a paradigm shift: pre-trained language models can serve as deployment-stable alternatives to synthetic imbalance correction in structured clinical prediction tasks. Beyond SSI, this finding underscores the potential of LLMs to improve the resilience of healthcare AI systems where minority-class prediction is critical.

Keywords: Imbalanced datasets; Artificial intelligence; Health informatics; Large language models; Machine learning; Minority class prediction; Mixture of Experts; Surgical site infection

***Corresponding author:**Stephen Russell
(russell@uwf.edu)

Citation: Russell S. Beyond SMOTE: Evaluating large language models and mixture of experts for prediction of surgical site infections. *Artif Intell Health*. 2026;3(3):025400082. doi: 10.36922/AIH025400082

Received: September 30, 2025**Revised:** October 23, 2025**Accepted:** October 24, 2025**Published online:** November 4, 2025

Copyright: © 2025 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Surgical site infections (SSIs) represent one of the most persistent and costly challenges in post-operative care, contributing to increased morbidity, mortality, and hospital readmission rates. In the United States, SSIs are responsible for up to \$10 billion in direct annual healthcare costs, with each infection extending hospitalization by 7–10 days on average and elevating patient mortality risk.^{1,2} From an organizational perspective, SSIs directly influence provider reimbursement, quality metrics, and accreditation status and are increasingly subject to regulatory reporting. Early detection and prevention of SSIs are paramount to improving surgical outcomes and care provisioning overall. Timely identification of an SSI allows clinicians to intervene sooner (for instance, with targeted antibiotics or surgical debridement), limiting the infection's severity and preventing further complications.

Early detection of SSIs remains difficult due to their latent and often subtle onset. Artificial intelligence (AI) based on machine learning (ML) models offers a promising solution, but they are hindered by two major barriers: (i) the sparsity of timely data features during clinical episodes³ and (ii) extreme or shifting class imbalance between SSI-positive and non-infected surgical cases in training and inference data.⁴ These two factors jointly reduce sensitivity and generalizability in deployed ML models, leading to elevated false positive rates or degraded performance under population shift.⁵ These factors have been shown to significantly impact the efficacy of ML predictive methods.⁶ The common remedy to the challenge of training data imbalance is the application of synthetic data augmentation such as Synthetic Minority Oversampling Technique SMOTE,⁷ over/undersampling, or hybrid methods to rebalance training data. However, these approaches often produce inconsistent performance under temporal distribution drift. In this study, we explore whether alternative ML modeling architectures, such as mixture of experts (MoE) and large language models (LLMs), can outperform traditional ML approaches and offer greater resilience to distribution drift. This work evaluates multiple methods for SSI prediction using structured electronic medical record (EMR) data, under varying target class imbalances, including an ensemble of classic ML models, an MoE framework, and a fine-tuned LLM. We evaluate these methods on real-world hospital data, including a temporally shifted evaluation cohort.

2. Data and methods

The objective of our research is to evaluate advanced ML strategies for handling extreme class imbalance in SSI prediction using structured EMR data. Traditional techniques such as SMOTE and undersampling are examined across ensemble classifiers, MoE, and LLM based on ModernBERT. The goal is to identify which approaches generalize best in terms of accuracy and robustness without relying on synthetic data or overly complex architectures. To better frame the contributions of our work in an interdisciplinary context, we provide some background from the literature.

2.1. Background and related work

2.1.1. Data imbalance correction in clinical ML

Class imbalance is widely acknowledged as a central challenge in clinical prediction.⁸ Common corrective methods include SMOTE, which generates new minority samples by interpolating existing ones; random over/undersampling, which reduces the majority class proportions; and class-weighted loss functions that penalize errors on the minority class more heavily. In

recent times, SMOTE has become quite popular. SMOTE attempts to balance class distribution in imbalanced datasets by generating synthetic examples for the minority class. It works by selecting a minority class sample and finding its k-nearest neighbors, then creating new synthetic data points along the line segments connecting the sample and its neighbors. This helps an ML classifier learn more generalized decision boundaries rather than overfitting to a few minority examples. Unlike over/undersampling, SMOTE⁷ introduces new, ideally plausible, examples rather than duplicating or eliminating existing ones.

Al-Ahmari and Nadeem⁹ found that applying SMOTE improved Matthews correlation coefficient (MCC) scores and yielded better balanced recall and precision for SSI prediction models. Other approaches, including over/undersampling and hybrid strategies, have also been employed but with mixed success depending on dataset characteristics.¹⁰ In addition to resampling techniques, improving the composition of the training dataset through better population selection has gained attention. Restricting modeling to specific surgery types (e.g., colorectal surgery^{11,12} and orthopedic procedures¹³) can help by increasing the prevalence of SSIs (positive target class samples) and improving model relevance. These studies indicate that restrictive sampling can yield better predictive ML performance by constraining training data.

SMOTE and hybrid strategies are widely used in clinical ML studies, but use of these approaches can suffer from several limitations: synthetic data may not preserve real clinical structure, semantics, and values;¹⁴ performance often degrades under test-time distribution shifts;¹⁵ and models may overfit synthetic or balanced training data but generalize poorly.¹⁶ Surprisingly, few studies systematically compare imbalance-aware model architectures under changing test distributions. Within the context of SSI prediction, even fewer explore architectures that are inherently robust to such imbalance without relying on synthetic sampling.

2.1.2. MoE and bidirectional encoder representations from transformers (BERT) LLM

While the focus of our study is not to make improvements in MoEs or LLMs directly, we provide a brief background on MoE and BERT LLMs here to illustrate the underlying methods. First introduced by Jacobs *et al.*,¹⁷ AI MoE is commonly used to enable efficient scaling, decomposition of complex problems, and computational savings in LLMs.¹⁸ To achieve these benefits, MoE architectures organize neural networks into a modular framework, in which several expert networks learn to specialize on different instances of the problem space while a separate

gating network determines which experts best contribute to a particular output. Each expert is an independent feed-forward network, and the system uses sparse activation so that only the most relevant experts are engaged for a given input.

An MoE “expert” does not necessarily denote specialized domain knowledge; instead, experts usually represent subnetworks tuned to handle certain data characteristics or functional tasks (e.g., reasoning or summarization).¹⁹ Figure 1 provides a schematic illustration: the gating network directs an input toward a subset of experts (e.g., Experts 1 and 3), applies weights to their outputs, and then combines them into a final prediction. In this study, we extend MoE to handle class imbalance by training experts on subsets of the data with different SSI/non-SSI proportions (e.g., 5/95%, 50/50%, 95/5%). Additional details on this are provided in the Methods section.

BERT is a foundational transformer-based LLM introduced by Devlin *et al.*²⁰ that enable deep bidirectional context understanding by jointly conditioning on both left and right word contexts in all layers.¹ Its architecture is based on the transformer encoder, allowing it to capture rich contextual relationships within text using self-attention mechanisms. Contemporary extensions of BERT, of which there are many, such as ModernBERT, improve upon BERT’s training efficiency and performance by incorporating architectural refinements like parameter sharing and sparse attention.²¹ These models have become commonplace in natural language processing tasks

ranging from question answering to clinical text extractive summarization due to their robust generalization and pretraining-transfer capabilities.²²

Recent advances in vision-language pre-training have further demonstrated the robustness and domain transfer potential of medical transformers. Approaches such as G2D,²³ Spectral Cross-Domain Neural Networks,²⁴ and parameter-efficient contrastive pre-training strategies like Freeze the Backbones²⁵ highlight the growing convergence between language- and vision-based medical AI models. These works support our motivation to explore pretrained transformer architectures for structured clinical data, where contextual embeddings can enhance robustness under distributional drift. Other recent works propose attention-based architectures that adapt transformers for tabular data, including TabTransformer²⁶ that applies attention to categorical feature embeddings while preserving independence across continuous features; FT-Transformer,²⁷ which treats all features as embeddings and applies full attention to model inter-feature interactions; and SAINT,²⁸ which adds inter-sample attention and data augmentation to improve generalization. While promising, these models have notable constraints, such as requiring custom tabular embeddings and limited support for free-text feature representations. We propose an alternative approach that re-casts structured tabular features into sentence sequences and fine-tunes a general-purpose transformer (ModernBERT) directly for classification. This approach allows us to use a pretrained model without architectural modification, avoids complex schema encoding, and any synthetic data generation.

2.1.3. SSI prediction

SSIs are a leading contributor to preventable postoperative morbidity and hospital readmissions. Despite their clinical and operational importance, early SSI detection remains difficult due to delayed symptom onset and complex etiology.^{1,2} Traditional SSI risk models rely on linear regression or scoring systems, often incorporating surgical type, comorbidities, and intra-operative variables.^{29,30} However, these models typically underperform in generalized clinical settings or lack interpretability when applied at scale. More recent work has leveraged ML to enhance prediction accuracy beyond traditional risk models.

Al Mamlook *et al.*³¹ used deep neural networks (DNNs) across 2.88 million surgical cases, reporting an area under the curve (AUC) > 0.85. Wu *et al.*³² developed an XGBoost-based model for arthroplasty-related SSIs with performance exceeding AUC 0.90. Many works in the literature show that incorporating observational (*i.e.*,

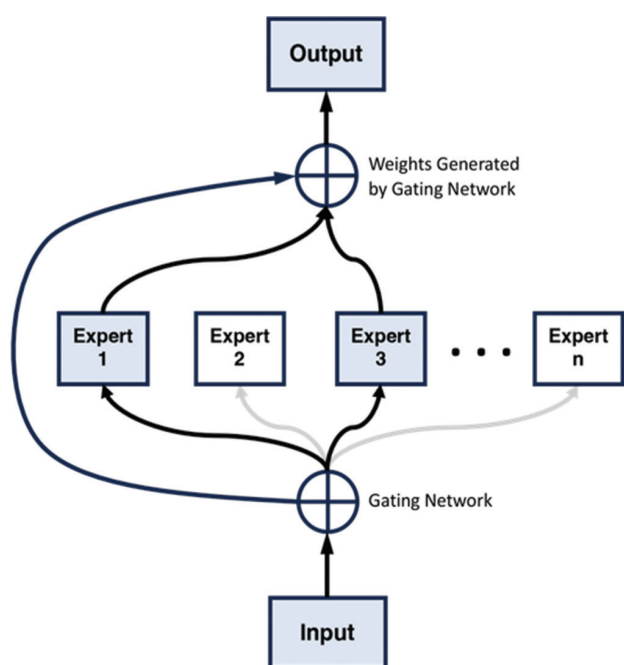


Figure 1. Basic mixture of experts architecture

bedside visuals and clinician notes) and clinical data (*i.e.*, labs/microbiology, imaging, and pharmacy) have a correlation to model performance. While such features are certainly an important consideration, one limitation in much of the prior work is the dependence on such observable and clinical features that frequently lag the SSI onset, are available only later in treatment timelines, or may not be available uniformly across facilities. Further, despite studies showing favorable results and accuracy, there were not many that gave emphasis on the occurrence of false positives.

A persistent challenge across many studies is the intrinsic severe class imbalance of SSI and non-SSI populations. SSIs typically occur in a relatively small percentage of surgical cases, often between 5% and 10% of surgical encounters,³¹ resulting in datasets heavily skewed toward the non-SSI-negative class. Without some form of correction, ML models will tend to favor predicting the majority class, yielding deceptively high accuracy but poor sensitivity and precision for true SSIs.³³ There is ample research that has shown that models trained without addressing imbalance suffer from low-positive predictive value despite acceptable AUCs.³⁴ Petrosyan *et al.*³⁵ reported an AUC of 0.89 using Random Forest but achieved a precision of only 34%, meaning two-thirds of predicted SSI cases were false positives. Xiong *et al.*³⁶ faced similar issues with an AdaBoost model for spinal surgery SSIs, achieving only 0.23 precision, and applied SMOTE⁷ to improve precision scores to 0.625 with only minor reduction in recall. Much of the prior work attempts to address the imbalance problem through over/undersampling or using synthetic data. Few studies explicitly benchmark how imbalance correction strategies perform under real-world shifts in data distribution. In contrast to the rich body of prior work, the objective of our research is to evaluate advanced ML strategies for handling extreme class imbalance in SSI prediction using structured EMR data.

2.2. Data source and population selection

The dataset analyzed in this study was de-identified before receipt by the investigators. As such, this work does not constitute human subjects research and did not require IRB approval under prevailing regulations. The study complied with all relevant guidelines for the use of de-identified health data. All data were collected passively and de-identified in compliance with the Health Insurance Portability and Accountability Act. Studies performed on de-identified data constitute non-human subject studies. Data access was governed by an internal data use agreement and is not publicly available. The initial dataset consisted of 151,733 inpatient cases (encounters) that are an inpatient census, spanning 2 years (2022 and 2023), drawn from a

large health care system's EMR system. Within the dataset, 20,062 (7.6%) were elective (versus emergency) surgeries; 73,300 (48%) patients were males; and there were 109,822 unique patients. Table S1 (in Supplementary File) provides a full list of features in the datasets. SSI-positive encounters are identified by SSI-related International Classification of Diseases version 10 (ICD-10) diagnosis (Dx) codes (*e.g.*, T81.4XXX) in the encounter final coding. ICD-10 diagnosis codes are standardized alphanumeric codes used internationally to classify and record diseases, symptoms, and other health conditions for clinical and billing purposes. A list of the relevant ICD-10 Dx codes utilized is presented in Table S2.

A specific class of SSI, those related to colon or hysterectomy procedures, was chosen to better align with the research in the literature and because SSIs related to these surgeries are of high importance to hospital quality metrics.³⁷ One might argue that the larger population could be filtered simply on encounters who had a colon or hysterectomy-related surgery and doing so would provide sufficient fidelity to restrict the majority class population. However, filtering on patients who had a colon or hysterectomy surgery assumes that all patients undergoing the same surgery are clinically comparable, but this is false. Two patients may undergo a colectomy but may differ significantly in many ways: comorbidities (*e.g.*, diabetes, immunosuppression, malnutrition); preoperative infection status; emergency versus elective context; and concurrent procedures or complications. These differences can affect SSI risk.³⁰ While potentially improving the majority class imbalance issues, filtering only on surgical type can aggregate dissimilar patients and introduce heterogeneity that can confuse an ML model during training. This type of qualitative or rule-based filtering to produce a representative dataset is typical in most ML training activities. However, such interpretive filtering is exceptionally sensitive to distribution changes resulting from context shifts, procedural or administrative adjustments (*e.g.*, treatment or medical coding changes), or other external conditioning.³⁸ With our research's focus on examining drift, adding an additional method to give more rigor to ensure both training and inference data are consistent, a method of determining clinical similarity is applied to further constrain the representativeness of both training and inference-input data. Of the census dataset encounters, 5,282 (3%) were colon and hysterectomy SSI positive.

To provide a method of ensuring clinically representative majority class selection, we employ a simple method of similarity using the full set of encounters' diagnoses and SSI-related procedure codes. Here, the set of diagnosis

codes and the set of procedure codes can be obtained, where the Jaccard distance J of two encounters results in a 2D vector of diagnosis and procedure code Jaccard distances and can be pairwise compared with minority class encounters using cosine similarity $C(\vec{v}_i, \vec{v}_j)$. Thus, encounter similarity S can be measured as a distance in the following manner: let $E_{SSI} = \{e_1, e_2, \dots, e_n\}$ be the set of reference SSI-positive encounters and $E_{neg} = \{e'_1, e'_2, \dots, e'_m\}$ be the non-SSI encounters. For each encounter pair $(e'_i, \dots, e'_j)$, compute:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (I)$$

$$C(\vec{v}_i, \vec{v}_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{\|\vec{v}_i\| \cdot \|\vec{v}_j\|} \quad (II)$$

$$\text{Sim}(e_i, e'_j) = C(J_{Dx}(e_i, e'_j), J_{PCS}(e_i, e'_j))^{[1, 1]} \quad (III)$$

A non-SSI case e'_j is retained if $\text{Sim}(e_i, e'_j) \geq 0.70$ for at least one $e_i \in E_{SSI}$. We set the similarity threshold at 0.70 to retain non-SSI encounters, guided by prior applications in clinical similarity modeling where thresholds between 0.65 and 0.75 are considered effective for balancing specificity and coverage. Similarity metrics such as the Jaccard and cosine distance have been widely used in patient similarity and clinical data matching tasks (e.g., in de-duplication of EMR notes using Jaccard at 0.7),³⁹ and in case-based reasoning systems achieving optimal sensitivity-accuracy trade-offs at the 0.70 threshold.⁴⁰ Thresholds much above 0.75 could exclude clinically analogous control exemplars, reducing cohort diversity, while lower thresholds (below 0.65) may inject heterogeneous majority-class contributors, diluting valid encounters. In addition, a manual review of selected sample pairs across thresholds indicated that a 0.70 threshold best preserved clinical plausibility and interpretability. Thus, our threshold represents an intentional but tunable compromise; one with reasonable interpretative clarity, transparency, and implementation-feasibility suitable for clinical deployments. The threshold value of 0.70 resulted in reducing the 151,733 records to 15,339; 5,282 inpatient SSI encounters (approximately 34%), and 10,057 “similar” non-SSI inpatient encounters were incorporated in a final dataset. Figure S1 provides a summarized diagram illustrating the EMR cohort selection and modeling pipeline.

2.3. Problem formulation and model configuration

We construct the problem as a tabular-data binary (SSI-negative/SSI positive: 0/1) classification problem subject to different training data target class proportions. In this regimen, several models are constructed. The baseline model configuration consists of classic ML models such as CatBoost,⁴¹ XGBoost, Random Forest, Light Gradient Boost

Machine, and Logistic Regression. These are examined with varying imbalances and configured standalone and in bias-wins/voting ensembles. We also adopt an MoE configuration with experts trained on incrementally proportioned target imbalance. Finally, a transformer-based LLM (ModernBERT) is examined, where the tabular input data are converted to a sentence-like structure for prediction of the binary class. Four configurations are evaluated (with and without employing SMOTE to address minority class imbalance): (i) best performing classical ML model; (ii) voting ensemble of classical ML models; (iii) MoE with experts trained with varied imbalance; and (iv) transformer-based LLM that predicts the binary target class.

2.3.1. MoE

The MoE configurations consisted of several “experts” each trained on differently proportioned imbalance, similar to ratios employed in the classical ML configurations (e.g., 5–95%, 10–90%, and 20–80%), as well as configurations with fewer numbers of experts, e.g., 5–95%, 50–50%, and 95–5%. One way to think of the MoE configuration is that experts trained with high imbalance are skewed to predict the biased target class and the gating network weights their performance in conjunction with other experts. Each expert $f_k(\cdot)$ is trained on a dataset with an imbalance ratio $r_k \in (0, 1]$. Let x be an input feature vector. The gating network $G(x)$ produces:

$$\alpha_k = \frac{r_k \cdot \exp(w_k^\top x + b_k)}{\sum_{j=1}^K r_j \cdot \exp(w_j^\top x + b_j)} \quad (IV)$$

$$f_{MoE}(x) = \sum_{k=1}^K \alpha_k f_k(x) \quad (V)$$

2.3.2. BERT LLM (ModernBERT)

The LLM classifier requires an additional pre-processing step and fine-tuning before inference. Each training record is transformed into a sentence-like sequence, including converting any numeric values to words, to create a human-readable value pair, such as this patient is 65 years old, is a male.: yes, where yes/no represents the binary target label, as illustrated in Figure 2. The model tokenizes the input into a sequence of tokens $x_1, x_2, \dots, x_n$, which is encoded as a hidden representation $h = \text{BERT}(x_1, x_2, \dots, x_n)$. The pooled embedding $h_{[CLS]}$, corresponding to the special classification token, is used for prediction through $\hat{y} = \text{sigmoid}(W^\top h_{[CLS]} + b)$. The binary cross-entropy loss is computed as $L_{BCE} = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$.

2.4. Model training and evaluation

All models consumed the same underlying EMR features. For ModernBERT, categorical values were rendered as

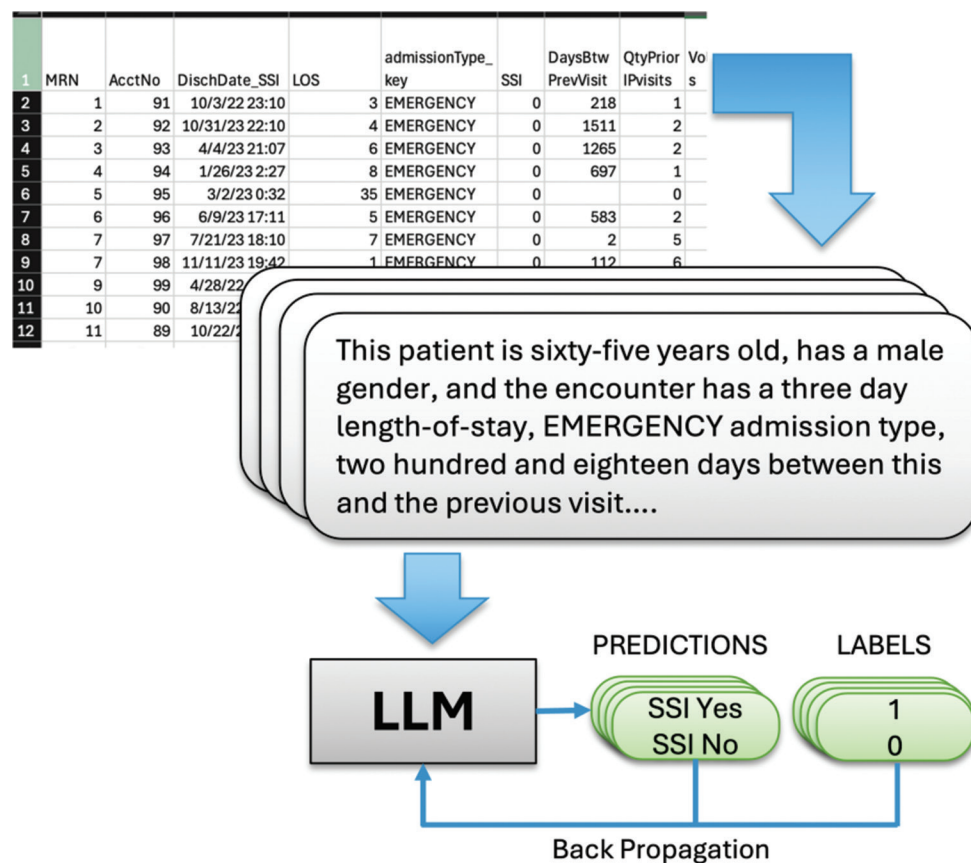


Figure 2. ModernBERT pre-processing and fine-tuning

human-readable tokens (e.g., “male,” “diabetic,” “emergency procedure”), numeric values were spelled out (e.g., “age sixty-five,” “body mass index of twenty-seven point three”), and missing values were imputed with the placeholder token [unknown] to maintain positional consistency. This produced a uniform sentence-like representation while preserving the original tabular information.

To construct the various proportioned imbalance datasets, before splitting into train-test sets, the 15,339-record dataset was randomly sampled, without replacement, to create sets of varying positive and negative target class proportions, e.g., 5–95%, 10–90%, 20–80%..., 90–10%, 95–5%, and the organic imbalance, 34–66%. The imbalanced sets were used for training and testing the models. To do so, the proportioned data were separated into two randomly selected subsets: (i) a training set and (ii) a hold-out test set, with an 80/20 train-test ratio. A 10-fold testing iteration was applied to assess the average results. A separate evaluation dataset, taken from 2024 encounters, was also collected and processed the same way as the initial 2022–2023 dataset, including diagnosis and procedural code similarity filtering. This 772-encounter evaluation dataset had an 8% minority class imbalance, unlike the 34% imbalance in the training dataset.

To examine the amount of distribution shift between the training data and the evaluation set, the numeric features were tested using a Kolmogorov–Smirnov (KS) test. 30% of the numeric features showed moderate to large shift (KS statistics between 0.10 and 0.25, with p -values below 0.01). The three most shifted numerical features were the target class indicator, pre-procedure hours, and surgical procedure volume. To further confirm the presence of a distribution shift and evaluate the shift in class-based features, we created a Random Forest classifier to predict whether data belonged to the training dataset or the evaluation set and employed a 70-20-10 training-test-validation split. The results of validation inference showed accuracy = 98.45%; precision=87.68%; recall = 78.57%; F1 score = 82.88%; MCC = 0.822; and cross-validated AUC = 0.948. The model’s training and test scores were similar.

Using the distribution-shifted evaluation dataset for inference, each configuration is compared across several metrics: accuracy, precision (selectivity), recall (sensitivity), AUC, F1 (a measure of the balance between precision and recall, based on the harmonic mean of both), and MCC. MCC is considered a better overall metric for

evaluating binary classification models than precision, recall, or F1 score, especially when dealing with imbalanced datasets.⁴² MCC considers all four components of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), as shown in Equation VI.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (VI)$$

For all models, the dataset was divided into training and validation subsets using an 80/20 split. For comparability, all models used identical splits and the same pre-processed features, outside of the additional value-to-text transformation noted above for the LLM training data. For training of the classical and ensemble models, multiple candidate algorithms including Logistic Regression, Random Forest, Gradient Boosting, and Light Gradient Boosting Machine were initialized with standardized preprocessing, one-hot encoding, and feature scaling applied. Model optimization employed five-fold cross-validation on the training data to identify the best-performing algorithm based on the MCC score. The best performing classical model, once identified, had hyperparameters (tree depth, learning rate, number of estimators) tuned through Tree-structured Parzen Estimator-based Bayesian optimization from the Optuna library.

The proportional DNNs were first trained as expert models for subsequent integration into the MoE architecture. Here, the global 80% training dataset split was further partitioned into 80% for training and an internal 20% portion (*i.e.*, 16% of the full dataset) reserved for validation during training and early stopping. Each expert was trained on a class-proportional subset to preserve the natural class-frequency distribution while maintaining balanced feature representation. Each expert consisted of three fully connected layers (512, 256, and 128 units) with ReLU activations and dropout (rate = 0.3) to mitigate overfitting. Training employed the AdamW optimizer (learning rate = 1e-4, weight decay = 0.01) for up to 50 epochs with a batch size of 128. Early stopping with a patience of three epochs was applied based on validation loss, and a cosine learning-rate scheduler was used to stabilize convergence. Experts were independently parameterized and no parameter sharing was applied. After convergence, expert weights were frozen and validated individually before integration into the MoE. The MoE architecture was constructed by combining the frozen proportional DNNs under a shallow feed-forward gating network (two hidden layers of 64 and 32 units with ReLU activations and dropout = 0.2). The gating network

received the same standardized feature inputs as the experts and was trained on the same 80% training split to assign softmax-normalized mixture weights $G(x)$ over the expert outputs, dynamically weighting their contributions during inference. Optimization used identical settings to the experts for consistency, with validation loss monitored for each epoch and checkpointing applied to preserve the best-performing weights. The final MoE model was then evaluated on the held-out 20% test set using accuracy, F1 score, precision, recall, and loss metrics to assess convergence behavior and generalization performance.

In fine-tuning the LLM, text tokenization was performed using a PreTrainedTokenizerFast/AutoTokenizer initialized from the ModernBERT-large checkpoint (vocabulary size = 50,368). The tokenizer applied truncation and padding to a maximum sequence length of 1,524 tokens. The average input length was 835 tokens with a standard deviation of 73 and maximum of 1,091; well within ModernBERT's input window. Model fine-tuning was implemented using the Hugging Face Trainer and TrainingArguments APIs for 10 epochs with a per-device batch size of 8. Optimization employed the AdamW optimizer (learning rate = 5e-5, weight decay = 0.01) with a linear learning-rate scheduler and 500 warmup steps. Early stopping was implemented through the Hugging Face EarlyStoppingCallback with a patience of two evaluation rounds. Evaluation was performed at the end of each epoch, and model performance was assessed using a custom compute_metrics function reporting the MCC, F1 score, precision, and recall.

3. Results

The baseline classic ML model performed consistently with or exceeding the state-of-the-art for SSI prediction, based on EMR features. Table 1 shows the results from the classic ML models on the initial (34%) imbalance dataset and SMOTE applied.

Figure 3 shows the MCC performance of the classic ML and DNN expert under varying levels of training data imbalance. The classic ML 50/50 imbalance configuration performed roughly the same, with and without SMOTE. The individual DNN experts did not perform as well as the classic ML. However, their results do not represent the MoE configuration, as no gating network has been employed at this stage.

3.1. Baseline model performance with distribution-shifted evaluation dataset

Table 2 shows how the final baseline models would have performed in production on the distribution-shifted evaluation dataset. The shifted distribution caused the

Table 1. Performance of classic machine learning models with SMOTE (34% SSI prevalence)

Model	Accuracy	AUC	Recall	Precision	F1 score	MCC
CatBoost	0.846	0.905	0.732	0.804	0.766	0.706
LightGBM	0.842	0.903	0.735	0.791	0.762	0.695
XGBoost	0.841	0.900	0.730	0.791	0.759	0.689
Random forest	0.820	0.882	0.731	0.742	0.736	0.652
Logistic regression	0.729	0.781	0.666	0.595	0.628	0.422

Abbreviations: AUC: Area under the curve; MCC: Matthews correlation coefficient; SSI: Surgical site infection; LightGBM: Light gradient boosting machine.

Table 2. Baseline performance on evaluation dataset (8% SSI prevalence)

Model	Accuracy	AUC	Recall	Precision	F1 score	MCC
CatBoost w/ SMOTE	0.859	0.894	0.328	0.551	0.411	0.351
DNN_50perc w/SMOTE	0.234	0.783	0.841	0.091	0.164	0.012

Abbreviations: AUC: Area under the curve; MCC: Matthews correlation coefficient; SSI: Surgical site infection.

seemingly viable CatBoost baseline model to have results that would be unusable in practice and an unexpected degradation in production, given the model's train, test, and validation performance.

3.2. Ensemble, MoE, and LLM performance with distribution-shifted dataset

Variations of each of the four categories of models, namely Classic ML, Ensemble, MoE, and LLM, have different proportional configurations. The configurations are as follows: CatBoost trained with baseline imbalance; ModernBERT LLM trained with baseline imbalance; Voting Ensemble with LLM and 50% imbalance CatBoost; Voting Ensemble with CatBoost 5%, 50%, and 95% SSI-positive imbalance; Ensemble Bias wins, 50% tie-break CatBoost 5% to 95% SSI-positive imbalance; CatBoost with 5% SSI-positive imbalance; CatBoost with 95% SSI-positive imbalance; MoE with 5% and 95% SSI-positive experts (LowHi); MoE with 5%, 50%, and 95% SSI-positive experts (LowMedHi); MoE with 30% to 70% SSI-positive experts; MoE with 10% and 90% SSI-positive experts (LowHiOffset); and MoE with all the imbalance experts (5–95% SSI positive). Figure 4 shows the MCC scores for all the models with and without SMOTE.

ModernBERT LLM without SMOTE achieves the highest MCC scores indicating robustness to distribution

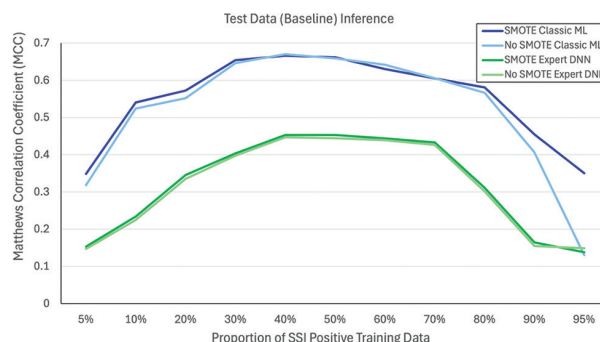


Figure 3. Classic ML and DNN experts with proportional imbalance
Abbreviations: DNN: Deep neural network; ML: Machine learning; SSI: Surgical site infection.

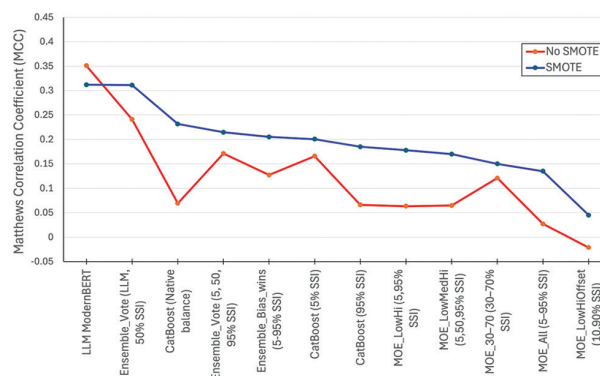


Figure 4. Model Matthews correlation coefficient scores for the distribution-shifted evaluation dataset

shift even without synthetic oversampling. Across the remaining models, SMOTE consistently improves performance, with a clear gap in MCC between SMOTE and no SMOTE variants, especially in Ensemble and MoE-based models. Figure 5 provides a heatmap comparing the performance of various model configurations with and without SMOTE across six metrics.

ModernBERT LLM consistently display superior performance in 5/6 metrics, especially without SMOTE. Figure 6 shows normalized scores for the best-performing model in each of the four model categories: LLM, Ensemble, CatBoost, and MoE. Normalizing the six scores, such that they all get equal weight, highlights that ModernBERT LLM with no SMOTE outperforms the other by 14% or more with the exception of recall, which is lower. However, ModernBERT LLM delivers a better balance of recall and precision over the others without the need for synthetic oversampling.

Ensemble_LLM_Mid (SMOTE) shows solid recall but lags significantly in precision, highlighting the well-known precision-recall trade-off effect. CatBoost Native

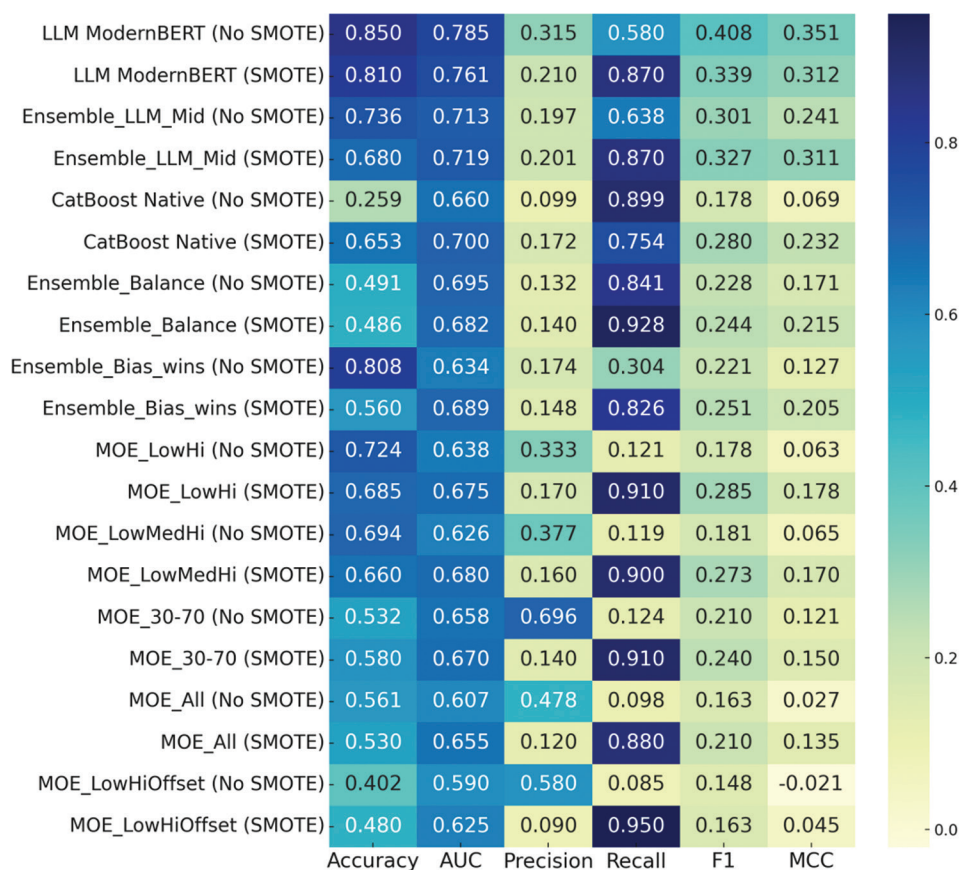


Figure 5. Model performance scores of SMOTE versus no-SMOTE

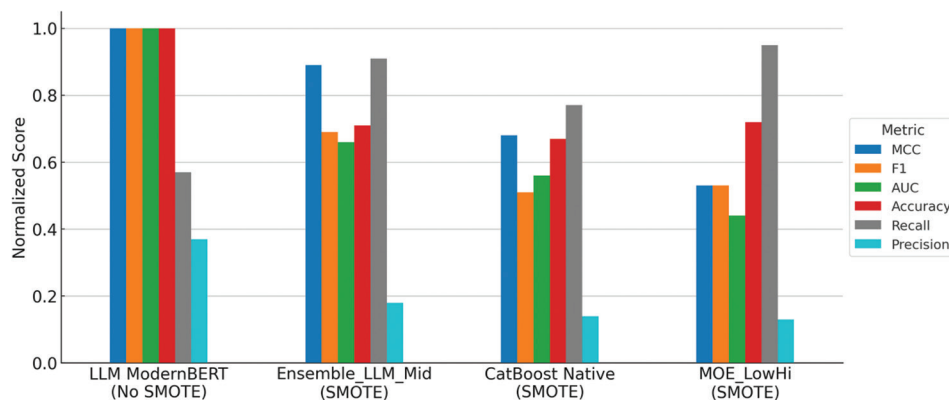


Figure 6. Normalized performance across the four model groups
Abbreviations: AUC: Area under the curve; MCC: Matthews correlation coefficient.

and MOE_LowHi with SMOTE show improved recall, but this comes at the expense of lower precision and overall metric balance compared to the LLM. Even with SMOTE, the best performing non-LLM models only close the gap in recall, not in precision, F1 score, or MCC; metrics all of which are better indicators of model robustness in binary classification of imbalanced data.

3.3. SSI prediction

Although slightly lower than the best performing models in the literature, the baseline CatBoost model's scores are consistent and competitive relative to most of the prior work. It is particularly noteworthy that the best performing models from the literature relied on a greater number and

more clinically specific features (such as laboratory results and/or observations such as swelling, fever, or purulent discharge); many of which are difficult to obtain in real (or close to real) time from EMR systems. Further, many published SSI prediction models lack additional validation and have only been tested retrospectively without distribution shift. The performance of ModernBERT LLM demonstrated superior generalization across distribution shifts without needing any form of data balancing.

4. Discussion

Prior works have shown that synthetic oversampling approaches such as SMOTE can generate clinically implausible examples and provide misleading performance gains during model development.^{14,15,43} Our results reinforce these concerns: while CatBoost with SMOTE performed well on balanced training and validation data, its performance degraded sharply under temporal distribution shift. This finding illustrates how reliance on synthetic balancing can create false confidence and underscores the risks of deploying resampled models in clinical practice.

Ensemble methods offered only modest robustness improvements. Bias-weighted voting and hybrid LLM ensembles reduced sensitivity to imbalance in some configurations, but the gains were marginal compared to single-model performance. These results suggest that while ensembles can smooth model variance, they do not address the fundamental vulnerability of resampled approaches to real-world distributional change.

Although theoretically attractive for learning from diverse imbalance ratios, MoE architectures consistently underperformed. The lack of robustness persisted even when experts were trained at both extremes and intermediate class proportions. This indicates that class-ratio specialization alone is insufficient for complex, heterogeneous EMR data. The underperformance of MoE models likely reflects limited sample size per expert and unstable gating under sparse activation. In imbalanced EMR regimes, experts may overfit narrow distributions while the gate receives weak gradients when class proportions diverge. Prior theory suggests MoE gains depend on both expert diversity and calibration; our results align with that view, indicating a need for stronger routing and confidence calibration. In their current form, MoE models are not a reliable solution for healthcare imbalance problems, and future advances will likely require more sophisticated gating mechanisms or confidence-calibrated routing.

The most significant outcome of our study was the superior and consistent performance of a fine-tuned ModernBERT model. Unlike models trained with SMOTE or resampling, the LLM generalized effectively across

temporal drift without artificial rebalancing. Several factors explain this resilience: transforming tabular features into natural language-like sequences allowed the model to exploit pretrained embeddings; attention mechanisms captured cross-feature dependencies often missed by tree-based methods; and the model avoided the overfitting tendencies of synthetic oversampling.

While these results are quite promising, they are not without some practical considerations. The robustness observed here for SSI prediction likely extends to other low-prevalence outcomes (*e.g.*, sepsis, adverse drug events, early readmission). Such tasks share extreme imbalance and heterogeneous feature interactions; pretrained language models provide a generalized path toward drift-resilient prediction without task-specific resampling. From a clinical deployment perspective, the fine-tuned ModernBERT model can be integrated into EMR-linked surveillance to flag higher-risk postoperative cases in near real time. Operating on routinely available structured EMR fields rather than delayed laboratory inputs, the model can augment infection-control workflows with early alerts while preserving transparency through feature-attribution overlays; interpretability remains essential for clinician trust. In practice, attention-weight visualizations and SHAP/Integrated Gradients can highlight factors (*e.g.*, peri-operative duration, wound class, emergency status) that most strongly influence predictions, enabling auditability and review in clinical dashboards.

Considerations regarding the underlying dataset employed in this study also frame the orientation and generalization of the results. For example, our dataset originates from multiple health systems, and it includes multiple hospitals and surgical specialties, offering a heterogeneous sample of patient populations. This internal diversity provides a partial buffer against overfitting to institutional idiosyncrasies. It could also be argued that both the underlying data and models, particularly the LLM, have implications for fairness and bias. Although demographic and procedural subgroups (*e.g.*, sex, surgery type, elective vs. emergency) were not explicitly rebalanced, ModernBERT embeddings capture cross-feature dependencies that may partially adjust for disparities. However, pretrained embeddings can also encode existing social or clinical biases. Although no significant subgroup-level disparities were observed during internal validation, future work will pursue multi-institutional validation to confirm generalizability and mitigate site-specific bias. We also plan to quantify fairness using subgroup-specific MCC and equality-of-opportunity metrics to assess differential error rates in the context of embedding-derived dependencies.

Taken together, our findings suggest that healthcare AI systems should move beyond reliance on synthetic rebalancing. Pretrained language models, when carefully adapted to structured clinical data, provide a deployment-stable path for minority class prediction. By demonstrating this in the context of SSI prediction, we provide evidence that LLM-based methods can improve robustness under drift and have broader potential across diverse clinical applications where class imbalance is the rule, rather than the exception.

5. Conclusion

Our study demonstrates that LLMs fine-tuned on tabular clinical data can replace traditional imbalance correction strategies, providing superior robustness under temporal and distributional drift. Unlike SMOTE and other resampling methods, which often inflate performance at training time but collapse under shift, ModernBERT generalized effectively to real-world SSI cohorts without artificial rebalancing. MoE architectures, while theoretically attractive, failed to provide meaningful robustness gains in practice. Ensemble configurations offered only marginal improvements. Together, these findings highlight a practical shift: rather than rebalancing data or increasing architectural complexity, adapting pretrained language models offers a more stable and deployment-ready path to reliable minority-class prediction in healthcare. From a practical perspective, our contributions are threefold:

- *Paradigm shift:* We provide early evidence that fine-tuned LLMs can serve as robust, deployment-stable alternatives to synthetic oversampling for imbalanced healthcare prediction tasks
- *Methodological guidance:* Similarity-based filtering improves organic class balance in clinical datasets, offering a simple, interpretable, and tunable pre-modeling step
- *Identified limitation:* MoE architectures, in their current form, are not yet reliable for clinical imbalance problems, underscoring the importance of deployment testing under drift.

Although this study relied on data from a single integrated health system, it encompassed multiple hospitals and surgical specialties, introducing operational heterogeneity that partially mitigates single-site bias. Formal multi-institutional external validation remains an important next step to confirm generalizability. Future work will also extend this approach to other emergent tabular LLMs (e.g., TabTransformer, Clinical ModernBERT, BioBERT) and across broader clinical contexts, with the goal of establishing generalizable standards for robust AI in healthcare.

Acknowledgments

None.

Funding

None.

Conflict of interest

The author declares no conflict of interest.

Author contributions

This is a single-authored paper.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

This data originates from multiple real-world health systems. Due to privacy and institutional regulations, the real-world healthcare data used in this study are not publicly available and cannot be shared. Assistance and insights on constructing similar data are available from the corresponding author upon reasonable request.

References

1. Leaper DJ, van Goor H, Reilly J, *et al.* Surgical site infection - a European perspective of incidence and economic burden. *Int Wound J.* 2004;1(4):247-273.
doi: 10.1111/j.1742-4801.2004.00067.x
2. Kirkland KB, Briggs JP, Trivette SL, Wilkinson WE, Sexton DJ. The impact of surgical-site infections in the 1990s: Attributable mortality, excess length of hospitalization, and extra costs. *Infect Control Hosp Epidemiol.* 1999;20(11):725-730.
doi: 10.1086/501572
3. Ghassemi M, Naumann T, Schulam P, Joshi AL, Suresh I, Nemati MR. A review of challenges and opportunities in machine learning for health. *AMIA Jt Summits Transl Sci Proc.* 2018;2018:191-200.
4. Salmi M, Atif D, Oliva D, Abraham A, Ventura S. Handling imbalanced medical datasets: Review of a decade of research. *Artif Intell Rev.* 2024;57(10):273.
doi: 10.1007/s10462-024-10884-2
5. Guo LL, Pföhl SR, Fries J, *et al.* Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Appl Clin Inform.* 2021;12(4):808-815.

- doi: 10.1055/s-0041-1735184
6. Fletcher RR, Olubeko O, Sonthalia H, *et al.* Application of Machine Learning to Prediction of Surgical Site Infection. In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE; 2019:2234-2237.
doi: 10.1109/embc.2019.8857942
7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321-357.
doi: 10.1613/jair.953
8. Luu J, Borisenko E, Przekop V, Patil A, Forrester JD, Choi J. Practical guide to building machine learning-based clinical prediction models using imbalanced datasets. *Trauma Surg Acute Care Open*. 2024;9(1):e001222.
doi:10.1136/tsaco-2023-001222
9. Al-Ahmari S, Nadeem F. Improving surgical site infection prediction using machine learning: Addressing challenges of highly imbalanced data. *Diagnostics (Basel)*. 2025;15:501.
doi: 10.3390/diagnostics15040501
10. Colborn KL, Bronsert M, Amioka E, Hammermeister K, Henderson WG, Meguid R. Identification of surgical site infections using electronic health record data. *Am J Infect Control*. 2018;46(11):1230-1235.
doi: 10.1016/j.ajic.2018.05.011
11. Cho SY, Kim YJ, Park J, *et al.* Development of machine learning models for the surveillance of colon surgical site infections. *J Hosp Infect*. 2024;146:224-231.
doi: 10.1016/j.jhin.2023.03.025
12. Shen Z, Chen Z, Zhang Y, *et al.* The development and validation of a novel model for predicting surgical complications in colorectal cancer of elderly patients: Results from 1008 cases. *Eur J Surg Oncol*. 2018;44(4):490-495.
doi: 10.1016/j.ejso.2017.12.013
13. Yeo I, Klemm C, Robinson MG, Esposito JG, Uzosike AC, Kwon YM. The use of artificial neural networks for the prediction of surgical site infection following TKA. *J Knee Surg*. 2022;36(06):637-643.
doi:10.1055/s-0041-1741396
14. Gholampour S. Impact of nature of medical data on machine and deep learning for imbalanced datasets: Clinical validity of SMOTE is questionable. *Mach Learn Knowl Extr*. 2024;6(2):827-841.
doi: 10.3390/make6020039
15. Sakho A, Malherbe E, Scornet E. Do we need rebalancing strategies? A theoretical and empirical study around SMOTE and its variants. *arXiv*. Preprint posted online 2024.
doi: 10.48550/arXiv.2402.03819
16. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc*. 2022;29(9):1525-1534.
doi: 10.1093/jamia/ocac093
17. Jacobs R, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. *Neural Comput*. 1991;3(1):79-87.
doi: 10.1162/neco.1991.3.1.79
18. Cai W, Jiang J, Wang F, Tang J, Kim S, Huang J. A Survey on Mixture of Experts. *TechRxiv*. Preprint posted online July 9, 2024.
doi: 10.36227/techrxiv.172055626.64129172/v1
19. Du H, Liu G, Lin Y, *et al.* Mixture of Experts for Intelligent Networks: A Large Language Model-enabled Approach. In: *2024 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE; 2024:531-536.
doi: 10.1109/iwcmc61514.2024.10592370
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*; 2019:4171-4186.
21. Warner B, Chaffin A, Clavié B, *et al.* Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv*. Preprint posted online 2024.
doi: 10.48550/arXiv.2412.13663
22. Lee J, Yoon W, Kim S, *et al.* BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.
doi: 10.1093/bioinformatics/btz682
23. Liu C, Ouyang C, Cheng S, Shah A, Bai W, Arcucci R. G2D: From Global to Dense Radiography Representation Learning Via Vision-Language Pre-Training. In: *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*; 2024.
24. Liu C, Cheng S, Ding W, Arcucci R. Spectral cross-domain neural network with soft-adaptive threshold spectral enhancement. *IEEE Trans Neural Netw Learning Syst*. 2025;36(1):692-703.
doi: 10.1109/tnnls.2023.3332217
25. Qin J, Liu C, Cheng S, Guo Y, Arcucci R. Freeze the Backbones: a Parameter-Efficient Contrastive Approach to Robust Medical Vision-Language Pre-Training. In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE; 2024:1686-1690.

- doi: 10.1109/icassp48485.2024.10447326
26. Huang X, Khetan A, Cvitkovic M, Karnin Z. TabTransformer: Tabular Data Modeling using Contextual Embeddings. In: *Proceedings of the 2020 Workshop on Tabular Representation Learning (NeurIPS 2020)*; 2020.
27. Gorishniy Y, Rubachev I, Khrulkov V, Babenko A. Revisiting deep learning models for tabular data. In: *2021 35th International Conference on Neural Information Processing System (NIPS)*. NIPS; 2021: 18932 - 18940 proceeding. 2021;34:18932-18943. Available on: https://proceedings.neurips.cc/paper_files/paper/2021/file/9d86d83f925f2149e9edb0ac3b49229c-Paper.pdf
28. Somepalli G, Goldblum M, Schwarzschild A, Bruss CB, Goldstein T. SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training. *arXiv*. Preprint posted online 2021.
doi: 10.48550/arXiv.2106.01342
29. van Walraven C, Musselman R. The surgical site infection risk score (SSIRS): A model to predict the risk of surgical site infections. *PLoS ONE*. 2013;8(6):e67167.
doi:10.1371/journal.pone.0067167
30. Korol E, Johnston K, Waser N, Sifakis F, Jafri HS, Lo M, Kyaw MH. A systematic review of risk factors associated with surgical site infections among surgical patients. *PLoS One*. 2013;8(12):e83743.
doi: 10.1371/journal.pone.0083743
31. Mamlook REA, Wells LJ, Sawyer R. Machine-learning models for predicting surgical site infections using patient pre-operative risk and surgical procedure factors. *Am J Infect Control*. 2023;51(5):544-550.
doi: 10.1016/j.ajic.2022.08.013
32. Wu G, Cheliger C, Southern DA, *et al*. Development of machine learning models for the detection of surgical site infections following total hip and knee arthroplasty: A multicenter cohort study. *Antimicrob Resist Infect Control*. 2023;12(1).
doi:10.1186/s13756-023-01294-0
33. van Boekel AM, van der Meijden SL, Arbous SM, *et al*. Systematic evaluation of machine learning models for postoperative surgical site infection prediction. *PLoS ONE*. 2024;19(12):e0312968.
doi: 10.1371/journal.pone.0312968
34. Heffernan A, Ganguli R, Sears I, Stephen AH, Heffernan DS. Choice of machine learning models is important to predict post-operative infections in surgical patients. *Surg Infect (Larchmt)*. 2025;26:520-529.
35. Petrosyan Y, Thavorn K, Smith G, *et al*. Predicting postoperative surgical site infection with administrative data: A random forests algorithm. *BMC Med Res Methodol*. 2021;21(1):179.
doi: 10.1089/sur.2024.288
- doi: 10.1186/s12874-021-01369-9
36. Xiong C, Zhao R, Xu J, *et al*. Construct and validate a predictive model for surgical site infection after posterior lumbar interbody fusion based on machine learning algorithm. *Comput Math Methods Med*. 2022;2022:2697841.
doi: 10.1155/2022/2697841
37. Sharafoddini A, Dubin JA, Lee J. Patient similarity in prediction models based on health data: A scoping review. *JMIR Med Inform*. 2017;5(1):e7.
doi: 10.2196/medinform.6730
38. Dockès J, Varoquaux G, Poline JB. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience*. 2021;10(9):giab055.
doi: 10.1093/gigascience/giab055
39. Gabriel RA, Kuo TT, McAuley J, Hsu CN. Identifying and characterizing highly similar notes in big clinical note datasets. *J Biomed Inform*. 2018;82:63-69.
doi: 10.1016/j.jbi.2018.04.009
40. Purwadi J, Delima R, Wibowo A, Rumuy A. Comparison of the application of weighted cosine similarity and Minkowski distance similarity methods in stroke diagnostic systems. *Int J Adv Comput Sci App*. 2023;14(12).
doi: 10.14569/ijacsa.2023.0141240
41. McElfresh DC, Khandagale S, Valverde JP, *et al*. When do Neural Nets Outperform Boosted Trees on Tabular Data? In: *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA; 2023:76336–76369.
42. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*. 2020;21(1):6.
doi: 10.1186/s12864-019-6413-7
43. Marchesi R, Micheletti N, Jurman G, Osmani V. Mitigating Health Data Poverty: Generative Approaches versus Resampling for Time-series Clinical Data. *arXiv*. Preprint posted online 2022.
doi: 10.48550/arXiv.2210.13958