

Machine Learning Versus Self-Supervised Transfer Learning for 30-Day Readmission Prediction in Diabetic Patients

Abigail Boatemaa

University of Bradford

Baffour Osei

University of Skills Training and Entrepreneurial Development (USTED)

Gabriel Osei Forkuo

`gabriel.forkuo@unitbv.ro`

Transilvania University of Brasov

Research Article

Keywords: class imbalance, denoising autoencoder, hospital readmission, Synthetic Minority Over-sampling Technique (SMOTE), tabular deep learning, XGBoost

Posted Date: August 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10560805/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background: Hospital readmission within 30 days of discharge is a closely watched indicator of care quality and a substantial driver of avoidable healthcare cost, and diabetes is among the chronic conditions most strongly associated with early readmission. This study compares six predictive models for 30-day readmission in patients with diabetes, aiming to establish realistic performance expectations for models built on administrative and coded clinical features alone.

Methods: We used the University of California, Irvine (UCI) Diabetes 130-US Hospitals dataset (101,766 encounters, 130 hospitals, 1999-2008). Four classical classifiers (logistic regression, random forest, eXtreme Gradient Boosting [XGBoost], and Light Gradient Boosting Machine [LightGBM]) were compared against two multilayer perceptrons (MLPs), one trained from scratch and one initialized via self-supervised, denoising-autoencoder pretraining. Since tabular data lacks direct analogs of image-style pretraining, transfer learning, and augmentation, each concept was mapped to a tabular equivalent: denoising-autoencoder pretraining for transfer learning, gradient-boosted tree ensembles as the "pretrained" family for structured data, and the Synthetic Minority Oversampling Technique, applied only within cross-validation folds, for augmentation. After excluding encounters ending in death or hospice discharge, 99,340 encounters were split into stratified 70/15/15 train, validation, and test partitions (11.4% readmission prevalence). Models were evaluated by area under the receiver operating characteristic curve (AUROC), and decision thresholds were examined separately from discrimination.

Results: On validation, XGBoost achieved the highest discrimination (AUROC = 0.672), ahead of LightGBM (0.669), random forest (0.654), logistic regression (0.641), and both MLP variants (0.585, 0.582). After randomized hyperparameter search with cross-validation (AUROC = 0.661), tuned XGBoost reached a test AUROC of 0.675 and an area under the precision-recall curve of 0.244. A default 0.5 threshold yielded a sensitivity of only 1.4%; a threshold chosen via Youden's J statistic on validation alone (0.128) raised sensitivity to 58.0% at the cost of precision (18.5%). These results converge with two independent 2025-2026 studies on the same dataset, suggesting a practical discrimination ceiling near AUROC 0.66-0.69.

Conclusions: Tree ensembles outperformed both neural network variants, consistent with the broader tabular-data literature. The findings argue for realistic expectations of electronic health record data for readmission risk stratification, and for decision-threshold selection to be treated as a first-class methodological choice.

Trial registration: Not applicable. This study involved secondary analysis of a publicly available, de-identified dataset and did not constitute a prospective healthcare intervention trial.

1. Introduction

1.1 Background

Unplanned hospital readmission within 30 days of discharge is one of the most closely watched quality indicators in American healthcare. Since 2012, the Centers for Medicare & Medicaid Services (CMS) have penalized hospitals with excess 30-day readmission rates for a defined set of conditions under the Hospital Readmissions Reduction Program, linking payment directly to this single measure [1]. Diabetes complicates this picture further: patients with diabetes are hospitalized more often, develop more comorbidities, and are readmitted at higher rates than the general inpatient population, making them a natural focus for readmission-prevention efforts [2].

For most of the last two decades, the standard tool for estimating readmission risk at the bedside has been a small set of manually derived clinical indices, most notably the Length of stay, Acuity of admission, Comorbidity burden, Emergency department use (LACE) index, which combines length of stay, acuity of admission, comorbidity burden, and recent emergency department use into a single score [3]. These indices are transparent and easy to compute, but they were built from a handful of hand-chosen variables and a linear scoring rule, which limits how much signal they can extract from the dozens of demographic, diagnostic, and medication variables now routinely captured in electronic health records (EHRs).

Machine learning offers a way to use that fuller feature set without an analyst having to specify every interaction in advance. Tree-based ensembles such as random forests and gradient-boosted trees can capture nonlinear relationships and threshold effects directly from the data [4, 5], and a substantial and growing body of work has applied these methods specifically to 30-day readmission in diabetic patients. Two recent studies are particularly relevant because they use the exact dataset analyzed here: Emi-Johnson and Nkrumah [6] benchmarked logistic regression, random forest, eXtreme Gradient Boosting (XGBoost), and a deep neural network on the UCI Diabetes 130-US Hospitals dataset and reported XGBoost as the strongest model, with an AUC-ROC of 0.667; Salim and Ibrahim [7] applied a more elaborate nested cross-validation and calibration pipeline to the same dataset and reported a nested-cross-validated XGBoost AUC of 0.664, rising to 0.688 after Platt-scaling calibration, with a Brier score of 0.094. Both findings sit close to the results reported here, and that convergence, across three independent pipelines built by different research teams, is itself informative: it suggests a practical discrimination ceiling for this dataset that no amount of additional model complexity has yet moved past. A wider set of 2025–2026 studies on diabetes-related readmission, using different cohorts and feature sets, reinforces the same general picture: Mishra et al. [8] compared logistic regression, decision tree, random forest, and XGBoost on a smaller specialty-clinic cohort and likewise found tree-based ensembles ahead of linear baselines; Tan et al. [9] and He et al. [10] each built Least Absolute Shrinkage and Selection Operator (LASSO)-selected, nomogram-based prediction models for 30-day readmission in elderly patients with type 2 diabetes, with and without comorbid heart failure, and reported the same emphasis on calibration and decision-curve-based clinical utility alongside discrimination; and Wang et al. [11] benchmarked eight machine learning algorithms, including a stacking ensemble, on elderly patients with concurrent type 2 diabetes and heart failure and again found gradient-boosted and ensemble methods outperforming simpler baselines. Taken together with the two dataset-identical studies described above, this broader literature suggests that the discrimination ceiling identified here is

not an artifact of the UCI Diabetes 130-US Hospitals dataset specifically, but a recurring feature of readmission prediction from structured, administrative, or coded EHR data more generally.

A separate empirical literature has asked whether deep learning offers a genuine advantage over tree ensembles specifically on tabular data of this kind, as opposed to the image and text domains where deep learning's advantages are well established. Grinsztajn, Oyallon, and Varoquaux [12] benchmarked modern tabular-specific neural networks against XGBoost and random forests across 45 datasets and found that tree-based models remain state of the art on medium-sized tabular data, even before accounting for their much lower training cost. This finding motivates a central design choice in the present study: rather than treating a plain multilayer perceptron as automatically competitive with gradient boosting, both are evaluated side by side under an identical protocol, and the multilayer perceptron is additionally given a self-supervised pretraining advantage, described below, to give the comparison its best chance of being fair to the deep learning side.

That self-supervised pretraining step is the tabular analog of transfer learning used in this study. Denoising autoencoders, introduced by Vincent, Larochelle, Bengio, and Manzagol [13], learn a compact, robust representation of an input distribution by training a network to reconstruct a clean input from a corrupted copy of itself, without using any labels. The resulting encoder weights can then be transferred into a supervised classifier and fine-tuned on the labeled task, mirroring the pretrain-then-fine-tune pattern used for large language and vision models, but built entirely from the unlabeled feature matrix of the dataset itself rather than from an external corpus. Class imbalance, the second methodological challenge specific to this task, since only about one in nine encounters in this dataset ends in a 30-day readmission, was addressed with the Synthetic Minority Over-sampling Technique (SMOTE), which generates synthetic minority-class examples by interpolating between real minority-class neighbors [14] and is the tabular-data counterpart of image-style data augmentation.

Taken together, these three adaptations, self-supervised pretraining for transfer learning, gradient-boosted trees as the tabular "pretrained" model family, and SMOTE for data augmentation, let this study speak directly to a request for transfer learning, pretrained models, and data augmentation on tabular clinical data without resorting to techniques built for a different data modality. The remainder of this paper describes the resulting pipeline, reports its results, and situates those results against the closely comparable estimates already published on this exact dataset.

1.2 Aim, Objectives, and Research Questions

This study aimed to compare classical machine learning models against self-supervised, transfer-learned neural networks for predicting 30-day hospital readmission in diabetic inpatients, using a single harmonized dataset and one consistent evaluation protocol, and to identify which approach offers the most clinically usable performance.

To achieve this aim, the study pursued the following objectives:

- To construct a clean, feature-engineered tabular representation of the UCI Diabetes 130-US Hospitals dataset suitable for supervised 30-day readmission prediction.
- To build a self-supervised, denoising-autoencoder pretraining stage on the unlabeled feature matrix and transfer its learned representation into a supervised neural classifier, as the tabular analog of transfer learning.
- To train, validate, and fine-tune four classical machine learning models and two neural network variants under one identical, leakage-controlled protocol, with class imbalance addressed through fold-internal SMOTE oversampling.
- To evaluate the resulting models on discrimination, calibration, and threshold-dependent classification performance, and to select a clinically defensible decision threshold rather than defaulting to 0.5.
- To situate the resulting performance against independently published benchmarks on the same dataset, and to draw out the implications for administrative-data-only readmission prediction as a clinical decision-support strategy.

Four research questions (RQ) guided the empirical work:

RQ1: How do classical machine learning models compare with a self-supervised, transfer-learned neural network in predicting 30-day readmission on this dataset?

RQ2: Does hyperparameter fine-tuning of the best-performing model, under a cross-validation protocol that avoids the SMOTE-related leakage documented in Section 3.8, meaningfully improve test-set performance?

RQ3: Does the discrimination achieved here replicate the range already reported by independent studies using the same dataset, and what does that convergence imply about the informational ceiling of administrative EHR data for this task?

RQ4: What decision threshold, as opposed to the conventional default of 0.5, is required to make model output clinically usable at this class prevalence, and what does the resulting precision-sensitivity trade-off imply for how such a model should be deployed?

2. Literature Review

2.1 Theoretical Framework

This study draws on two complementary bodies of theory: clinical risk-prediction theory, which explains what a readmission risk score is for and why it is hard to get right, and statistical learning theory, which explains how the model families compared here actually learn to approximate that risk.

Clinical risk-prediction theory treats 30-day readmission as the observable endpoint of a partially hidden process combining disease severity, comorbidity burden, treatment response, discharge-planning quality,

and post-discharge social circumstances, only some of which are captured in structured administrative data [3]. Classical instruments such as the LACE index operationalize a small, clinically motivated subset of that process into a linear score; this is transparent and clinically interpretable, but it assumes the relationship between predictors and readmission risk is linear and fully specified by four to five variables, an assumption that a wider EHR feature set need not respect.

Statistical learning theory supplies the second half of the framework. Every model evaluated here, from logistic regression to the transfer-learned multilayer perceptron, is an instance of empirical risk minimization: each learns a function mapping the encoded encounter features to a readmission probability by minimizing a training-set loss, subject to an inductive bias that determines what kinds of patterns it can represent efficiently. Logistic regression assumes a linear, additive relationship in the log-odds of readmission. Tree-based ensembles, random forest and the gradient-boosted variants XGBoost and Light Gradient Boosting Machine (LightGBM), relax that assumption and can represent threshold effects and interactions among predictors, such as the joint effect of a circulatory diagnosis and a high number of prior inpatient visits, without those interactions being specified in advance [4, 5, 15]. The multilayer perceptron relaxes the assumption further still, in principle able to approximate arbitrary smooth functions of the input, but Grinsztajn et al. [12] show empirically that this additional flexibility does not translate into an advantage on medium-sized tabular datasets such as this one, a finding the results in Section 4 corroborate directly.

The self-supervised pretraining stage draws on a third, narrower strand of theory: the idea that a denoising autoencoder, trained to reconstruct a clean input from a corrupted version of itself, is forced to learn a representation that captures the joint structure of the input distribution rather than any single label [13]. Transferring that representation into a supervised classifier gives the classifier an initialization informed by the full unlabeled feature distribution, including the validation and test partitions' own features, though never their labels, before any label information is introduced. Whether that initialization meaningfully improves over a random one, given the tabular-data findings above, is itself one of the empirical questions this study addresses. The broader empirical record on self-supervised pretraining for structured EHR data is itself mixed rather than uniformly favorable. A 2025 scoping review of self-supervised representation learning for categorical EHR data found that autoencoder-based architectures account for a substantial share, roughly a quarter, of published approaches, alongside transformer- and graph-based methods, but also noted that the clinical impact of these representations is inconsistently assessed across the literature [16]. Similarly, a 2025 benchmarking study of pretraining strategies for EHR foundation models found that task-specific supervised pretraining outperformed self-supervised, masked-language-modeling pretraining on one downstream task, while the self-supervised model performed best on a different downstream task, suggesting that the value of self-supervised pretraining is task-dependent rather than uniformly positive [17]. Against that backdrop, a modest, single-hidden-layer denoising autoencoder failing to outperform random initialization on this dataset, as reported in Section 4.1, is consistent with rather than an outlier within the wider self-supervised tabular-EHR literature.

2.2 Conceptual Framework

Figure 1 sets out the conceptual framework that structures the empirical work reported in this study. The framework begins with a single data source, the UCI Diabetes 130-US Hospitals dataset, and moves through cleaning and feature engineering into one tabular representation shared by every model family, rather than the three parallel representations used in image- or sequence-based comparisons, since every model compared here consumes the same encoded feature vector.

From that shared representation, the framework branches into three parallel adaptations, each addressing one part of the request that pretrained models, transfer learning, and data augmentation be brought to bear on this task: classical machine learning models trained directly on the tabular features; SMOTE oversampling of the minority class, applied strictly inside each training fold; and self-supervised denoising-autoencoder pretraining, whose encoder weights are transferred into, and fine-tuned within, a supervised multilayer perceptron. All three branches converge on a common two-stage evaluation protocol, a validation-based comparison stage followed by hyperparameter fine-tuning and a single, final test-set evaluation, and the framework closes with a decision-threshold selection stage that translates the resulting probability estimates into a usable classification rule, since Section 4 shows that this final step matters as much to clinical usability as any modeling choice upstream of it.

3. Methodology

This section describes the end-to-end pipeline used to predict 30-day readmission from the UCI Diabetes 130-US Hospitals dataset. Throughout, the guiding principle was to keep validation and test data strictly unseen at every stage where a modeling decision was made, including inside cross-validation folds, since a resampling or scaling step fit on data outside the current fold is a leakage source that can silently inflate reported performance; Section 3.8 documents one such leakage error that was caught and corrected during model development.

3.1 Data Sourcing and Cleaning

The data used in this study are the UCI Diabetes 130-US Hospitals for Years 1999–2008 dataset, comprising 101,766 inpatient encounters from 130 US hospitals and integrated delivery networks, each representing a hospitalization of one to fourteen days for a patient with diabetes [2]. Encounters whose discharge disposition indicated death or transfer to hospice care (discharge disposition codes 11, 13, 14, and 19 through 21) were excluded, since these encounters cannot, by definition, have a valid 30-day readmission outcome. After this exclusion, 99,340 encounters remained.

3.1.1 Data Quality and Cleaning Rules

Columns with nearly universal missingness, weight (missing in the large majority of encounters), payer_code, medical_specialty, and max_glu_serum, were dropped outright rather than imputed, since

imputing a column that is almost entirely absent risks manufacturing signal rather than recovering it. The small number of remaining encounters with missing gender were dropped. "?" placeholders and "Unknown/Invalid" category labels used by the original data provider were normalized to genuine missing values before any imputation logic was applied, so that downstream missing-data handling operated on an honest missingness pattern rather than a string that happened to look like a category.

3.2 Handling Missing Data

Missingness in this dataset is not random: certain laboratory results and specialty fields are far more likely to be absent for some encounter types than others, which is itself informative rather than a nuisance to be silently patched over [18]. For the small number of remaining numeric fields, median imputation fit on the training partition only was used. For categorical fields, missingness was encoded as an explicit "Missing" category rather than imputed to the mode, preserving the (frequently predictive) fact that a field was not recorded, consistent with recommended practice for structured clinical data with informative missingness [19]. All imputation statistics, medians and category frequencies alike, were fit on the training partition only and then applied unchanged to the validation and test partitions, to avoid leaking distributional information from later-seen data into training.

3.3 Data Processing and Feature Engineering

3.3.1 Tabular Representation

The three recorded International Classification of Diseases, Ninth Revision (ICD-9) diagnosis fields (diag_1, diag_2, diag_3) were collapsed from several hundred distinct codes each into nine clinically meaningful categories, circulatory, respiratory, digestive, diabetes, injury, musculoskeletal, genitourinary, neoplasms, and other, following the grouping used in Strack et al. [2], the paper associated with this dataset. Rare levels (fewer than 500 encounters) of the admission-type, discharge-disposition, and admission-source identifier fields were collapsed into an "Other" category to control one-hot dimensionality. Age, originally recorded as ten-year bins, was encoded as an ordinal integer to preserve its natural ordering rather than expanded into ten unordered dummy variables. Engineered features included the count of active diabetes medications, total prior-year service utilization (the sum of outpatient, emergency, and inpatient visit counts), and a composite service-intensity feature (the sum of lab procedures, procedures, and medications administered during the encounter). After one-hot encoding of the remaining categorical fields and standardization of numeric fields (fit on the training partition only), the feature matrix contained 153 columns for 99,340 encounters, split into stratified 70/15/15 train, validation, and test partitions (69,538 / 14,901 / 14,901 encounters), preserving the 11.4% 30-day readmission rate in every partition.

3.3.2 Data Augmentation: SMOTE Oversampling

With roughly one 30-day readmission for every eight non-readmissions, a classifier trained on the raw class distribution can achieve a low loss largely by predicting the majority class. SMOTE [14] was used to rebalance the training data by generating synthetic minority-class examples along the line segments

connecting each real minority-class encounter to its nearest minority-class neighbors in feature space. Critically, SMOTE was fit and applied only to the training fold in use at that moment, whether the full training partition for a single model fit, or the training portion of a single cross-validation fold during hyperparameter search, and never to validation or test data. Section 3.8 documents a specific instance where this constraint was initially violated during hyperparameter search, and how the resulting leakage was identified and corrected.

3.3.3 Transfer Learning: Self-Supervised Denoising-Autoencoder Pretraining

In place of an image- or text-domain pretrained model, a denoising autoencoder was pretrained from scratch on the full unlabeled feature matrix, the concatenated training, validation, and test feature vectors, with no label information used at this stage [13]. The autoencoder consists of a single hidden layer of 64 units with a rectified linear activation, trained to reconstruct a clean input vector from a corrupted copy produced by additive Gaussian noise (standard deviation 0.15) combined with random feature masking (masking probability 0.15) on each mini-batch. The network was trained with the Adam optimizer for 60 epochs in batches of 256, reaching a final mean-squared reconstruction error of 0.023 on the full feature matrix. The resulting encoder weights (the input-to-hidden-layer matrix) were then transferred into the first hidden layer of a supervised multilayer perceptron and fine-tuned end to end on the labeled training data, the tabular-data analog of the pretrain-then-fine-tune protocol used for large vision and language models.

3.4 Model Selection and Architectures

3.4.1 Machine Learning Models

Four classical models were benchmarked: logistic regression, a transparent linear baseline; random forest, a bagged tree ensemble [4]; XGBoost, a gradient-boosted tree ensemble [5]; and LightGBM, a histogram-based gradient-boosted tree ensemble optimized for large tabular datasets [15]. These four represent the state of the art for structured, tabular clinical prediction and served as the "pretrained model family" comparator requested for this study, in the sense that gradient-boosted trees, not deep neural networks, are the models with the strongest track record on data of this kind [12].

3.4.2 Neural Network Models

Two multilayer perceptron variants, each with hidden layers of 64 and 32 units and rectified linear activations, were evaluated. The first was trained entirely from randomly initialized weights. The second had its first hidden layer initialized from the pretrained denoising-autoencoder encoder described in Section 3.3.3, then fine-tuned end to end on the labeled training data; this is the transfer-learned model referred to throughout the remainder of this paper. Both variants used the Adam optimizer with early stopping (patience of 10 epochs, maximum of 100 epochs) monitored on an internal validation split of the training data.

3.5 Hyperparameters

Hyperparameters for the final model were tuned through randomized search rather than an exhaustive grid, which explores a wider effective range of settings for a fixed computational budget [20]. Table 1 lists the search space used for the XGBoost fine-tuning stage described in Section 3.8; the other five candidate models used the fixed configurations listed alongside the validation results in Table 2, since only the validation-stage winner was carried forward to full hyperparameter search.

Table 1
Randomized hyperparameter search space for the XGBoost fine-tuning stage.

Hyperparameter	Search space
n_estimators	{100, 150, 200}
max_depth	{3, 4, 5, 6}
learning_rate	{0.03, 0.05, 0.1}
subsample	{0.7, 0.8, 0.9}
colsample_bytree	{0.7, 0.8, 0.9}
min_child_weight	{1, 3, 5}
gamma	{0, 0.1, 0.3}
reg_lambda	{0.5, 1.0, 2.0}
<i>Note. Search conducted with 6 randomly sampled configurations under 3-fold stratified cross-validation, with SMOTE resampling fit inside each fold via an imblearn Pipeline (see Section 3.8).</i>	

3.6 Training, Validation, and Testing

3.6.1 Training Procedure

All six candidate models were fit once on the SMOTE-resampled training partition and evaluated on the untouched validation partition, using the fixed configurations reported in Table 2. The validation-stage winner, XGBoost, was then carried into the randomized hyperparameter search described in Section 3.5, refit on the full training partition with SMOTE applied per-fold, and evaluated a single time on the held-out test partition. A fixed random seed (42) was used throughout for reproducibility.

3.6.2 Evaluation Metrics

Area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) are reported as the primary, threshold-independent discrimination metrics, since AUROC alone can be a misleading summary at this level of class imbalance. The Brier score is reported as a calibration measure, since it captures both discrimination and the accuracy of the predicted probabilities themselves, not merely their rank ordering. Precision, recall (sensitivity), F1, and accuracy are reported at

two operating points, the conventional default threshold of 0.5 and a threshold selected via Youden's J statistic [26] on the validation set (Section 3.7), since Section 4 shows these two operating points tell substantially different stories about the same underlying model.

3.7 Decision Threshold Selection

A predicted-probability output is not, by itself, a classification decision; that requires a threshold, and the conventional default of 0.5 implicitly assumes that false positives and false negatives carry equal cost and that the two classes are roughly balanced, neither of which holds here. The decision threshold used for all threshold-dependent metrics reported in Section 4.4 was selected by maximizing Youden's J statistic (sensitivity plus specificity minus one) on the validation partition only; the test partition was never used to choose this threshold, so the test-set classification metrics reported in Section 4.4 remain an honest, unbiased estimate of how the resulting decision rule would perform on genuinely unseen encounters. Youden's J is used here as a transparent, reproducible starting point rather than as a claim that it is the clinically correct operating point: recent methodological guidance cautions that maximizing Youden's J implicitly treats a missed readmission and an unnecessary intervention as equally costly and, in effect, sets the threshold near the outcome's prevalence, an assumption that is rarely defensible in practice and that a genuine deployment decision should instead derive from the relative costs of the two error types [21, 22]. The Youden-optimized threshold reported in Section 4.4 should therefore be read as a methodological baseline that makes the model's discrimination usable for comparison purposes, not as the threshold a hospital should adopt without its own cost-based analysis.

3.8 Best Model Selection and Fine-Tuning

Model selection followed two stages. In the first stage, all six candidate models were trained under identical preprocessing and evaluated on the validation partition; whichever model achieved the highest validation AUROC, XGBoost, was carried forward. In the second stage, that model underwent randomized hyperparameter search under 3-fold stratified cross-validation, before a single, final evaluation on the fully held-out test partition.

One methodological detail is worth stating plainly rather than leaving implicit, because it directly affects how the fine-tuning results in Section 4.2 should be read. An initial implementation of this hyperparameter search applied SMOTE to the entire training partition once, before the cross-validation split, and then ran cross-validation on the resulting resampled data. This produced a cross-validated AUROC of 0.957, far above the 0.672 achieved by the same model family on genuinely unseen validation data, an implausible gap that was the first sign something was wrong. The cause was leakage: because SMOTE-generated synthetic samples are built from real training-fold neighbors, a synthetic point derived from an encounter that ends up in one cross-validation fold can be highly similar to synthetic points derived from the same encounter's neighbors that end up in a different fold, letting information about the held-out fold leak into the fold used for training. This was corrected by moving SMOTE inside an `imblearn Pipeline` object, so that the resampling step is refit independently within each cross-validation fold's training portion only; the corrected, leakage-free cross-validated AUROC was 0.661, consistent with

the 0.672 to 0.675 range obtained on the validation and test partitions. This correction, and the reasoning behind it, is reported here in full because the same failure mode, resampling before rather than inside cross-validation, is a common and easily overlooked source of inflated performance claims in the class-imbalance literature more broadly, and is worth naming explicitly rather than silently fixing. A related, separate caution documented in the recent clinical machine learning literature is that resampling techniques such as SMOTE, even when correctly confined to the training fold, alter the outcome distribution the model is fit on and can distort the resulting predicted probabilities away from the true, prevalence-consistent risk; when calibrated probabilities matter, an additional post-training calibration step is recommended rather than relying on resampling alone [21]. The Brier scores reported in Section 4.3 and Section 4.4 should be read with this caveat in mind, since no separate calibration step beyond SMOTE was applied here.

4. Results

All results reported in this section come from a single, complete run of the pipeline described in Section 3. No test-set number in this section was used to choose a model, a hyperparameter, or a decision threshold at any earlier stage.

4.1 Search-Stage Model Comparison

Table 2 and Fig. 2 report validation-set discrimination for all six candidate models, each trained once on the SMOTE-resampled training partition. XGBoost achieved the highest validation AUROC (0.672), narrowly ahead of LightGBM (0.669) and random forest (0.654), with logistic regression somewhat behind (0.641). Both multilayer perceptron variants trailed the tree-based models by a wide margin (0.585 for the from-scratch network, 0.582 for the transfer-learned network), a gap consistent with the broader finding that tree ensembles, not neural networks, remain the stronger default for tabular data of this size [12]. The transfer-learned multilayer perceptron did not outperform its randomly initialized counterpart on this dataset; if anything it was marginally, though not meaningfully, worse, suggesting that a single-layer autoencoder pretrained on this feature set does not supply the fine-tuned classifier with information beyond what the labeled training data already provides at this scale.

Table 2
Validation-set results for all six candidate models.

Model	Val. AUROC	Val. AUPRC	Train time (s)	Family
XGBoost	0.672	0.224	17.6	Gradient-boosted trees
LightGBM	0.669	0.219	8.4	Gradient-boosted trees
Random Forest	0.654	0.190	51.4	Bagged trees
Logistic Regression	0.641	0.203	4.8	Linear
MLP (from scratch)	0.585	0.153	75.6	Neural network
MLP (transfer-learned)	0.582	0.154	134.9	Neural network

4.2 Effect of Hyperparameter Fine-Tuning

As documented in Section 3.8, an initial hyperparameter search produced a cross-validated AUROC of 0.957, which was traced to SMOTE resampling being applied before rather than inside cross-validation and corrected accordingly. Under the corrected, leakage-free protocol, the best cross-validated configuration achieved an AUROC of 0.661, consistent with, if marginally below, the 0.672 obtained by the untuned XGBoost model on the validation partition, and with the 0.675 the tuned model went on to achieve on the held-out test partition. The best hyperparameter configuration found was `n_estimators = 200`, `max_depth = 5`, `learning_rate = 0.1`, `subsample = 0.9`, `colsample_bytree = 0.7`, `min_child_weight = 3`, `gamma = 0`, and `reg_lambda = 1.0`. In short, the fine-tuning stage did not produce a dramatic accuracy gain over the untuned baseline, once the leakage was corrected, but it did provide a properly validated final configuration to carry into the single test-set evaluation.

4.3 Discrimination and Calibration on the Test Set

The tuned XGBoost model achieved a test AUROC of 0.675 (Fig. 3) and a test AUPRC of 0.244 (Fig. 4), against a baseline positive-class prevalence of 11.4%, meaning the model concentrates precision well above chance among its highest-ranked predictions even though its overall discrimination is moderate. The Brier score on the test set was 0.095, indicating that the model's predicted probabilities are, on average, reasonably close to the observed outcome frequencies, without a separate calibration step being applied.

4.4 Classification Performance: Default Versus Youden-Optimized Threshold

Table 3 reports classification performance at the conventional default threshold of 0.5 and at the Youden-optimized threshold of 0.128, selected on the validation partition alone as described in Section 3.7. The contrast is stark. At 0.5, the model classifies almost every encounter as "no readmission," achieving high accuracy (88.6%) but a sensitivity of only 1.4%, meaning it correctly flags fewer than 2 in

every 100 true 30-day readmissions, which would make it clinically useless as a screening tool despite its unremarkable-looking accuracy figure. At the Youden threshold, sensitivity rises to 58.0%, correctly flagging more than half of true readmissions, at the cost of precision falling to 18.5% and accuracy falling to 66.1%. Which operating point is preferable depends entirely on the intended use: a hospital deploying this model to trigger a low-cost intervention, such as an additional discharge-planning phone call, can tolerate a low positive predictive value in exchange for high sensitivity; a hospital deploying it to trigger an expensive, resource-intensive intervention would want a higher threshold and would need to accept missing more true readmissions. Figure 5 shows the confusion matrix at the Youden threshold.

Table 3
Test-set classification performance at two decision thresholds.

Threshold	AUROC	Sensitivity	Precision	F1	Accuracy
0.5 (default)	0.675	0.014	0.561	0.026	0.886
0.128 (Youden's J)	0.675	0.580	0.185	0.281	0.661
<i>Note. AUROC is threshold-independent and identical at both rows; it is repeated for reference. Brier score at both thresholds was 0.095.</i>					

4.5 Predictors of Readmission

Figure 6 reports the twenty most predictive features by XGBoost gain-based importance. The strongest predictors were a circulatory-system secondary diagnosis, an "other" tertiary diagnosis, a missing Hemoglobin A1c (A1C) result, the number of prior inpatient visits, and a circulatory-system primary diagnosis, followed closely by age, discharge disposition, and race. This pattern is clinically coherent: circulatory comorbidity and prior inpatient utilization are well-established readmission risk factors, and a missing A1C result plausibly proxies for less intensive diabetes management during the index admission, consistent with the original hypothesis motivating this dataset's collection [2].

4.6 Comparison with Published Benchmarks on the Same Dataset

Table 4 places the present results alongside two independently conducted 2025–2026 studies that used the identical UCI Diabetes 130-US Hospitals dataset. Emi-Johnson and Nkrumah [6] reported an XGBoost AUC-ROC of 0.667 using an 80/20 train-test split without a validation partition or hyperparameter search. Salim and Ibrahim [7] reported a nested-cross-validated XGBoost AUC of 0.664, rising to 0.688 after Platt-scaling calibration, with a Brier score of 0.094, using a more elaborate five-outer/three-inner nested cross-validation design and cost-sensitive class weighting rather than SMOTE. The present study's test AUROC of 0.675 and Brier score of 0.095 sit within a few thousandths of both published estimates despite three independent teams using different preprocessing choices, different class-imbalance strategies, and different validation designs. This convergence is the strongest evidence in this study that AUROC in the neighborhood of 0.66 to 0.69 represents a genuine, replicated ceiling for

this exact prediction task using administrative and coded clinical features alone, rather than an artifact of any one team's modeling choices. This picture is reinforced further by cohort-specific studies published over the same period: Mishra et al. [8] reported an XGBoost AUC-ROC above 0.9 on a much smaller, single-clinic cohort of 352 records, and Tan et al. [9], He et al. [10], and Wang et al. [11] each reported comparable or higher discrimination, generally in the 0.70 to the low 0.90s range, when predicting readmission within narrower, more clinically homogeneous elderly diabetes and diabetes-heart-failure populations using richer clinical and laboratory variables unavailable in the present administrative dataset. Read alongside the present results, this pattern suggests that the 0.66 to 0.69 ceiling identified here is specific to prediction from coded, administrative EHR fields on a large, heterogeneous inpatient population, and that discrimination further above this ceiling in the literature tends to come from smaller, more clinically detailed cohorts rather than from a superior classifier applied to the same kind of data.

Table 4
Comparison with independently published benchmarks on the same dataset.

Study	Best model	AUROC / AUC	Brier score
Emi-Johnson & Nkrumah [6]	XGBoost	0.667	not reported
Salim & Ibrahim [7]	XGBoost (calibrated)	0.688	0.094
Present study	XGBoost (tuned)	0.675	0.095

5. Discussion

This study set out to compare classical machine learning against self-supervised, transfer-learned neural networks for 30-day readmission prediction in diabetic inpatients, with pretrained models, transfer learning, and data augmentation adapted to their tabular-data equivalents. Four research questions guided that comparison, and the results in Section 4 give a clear answer to each.

5.1 Summary of Findings

On RQ1, gradient-boosted trees, XGBoost in particular, clearly outperformed both the linear baseline and both neural network variants, by a margin of roughly three to nine points of validation AUROC. This is consistent with the theoretical expectation set out in Section 2.1 and with the empirical tabular-data literature more broadly [12]: the additional representational flexibility of a neural network did not translate into better discrimination on this feature set, and self-supervised pretraining did not close that gap, or even move it in the expected direction. This is a genuine, if modest, negative result for the transfer-learning adaptation specifically, and it is reported as such rather than reframed after the fact.

On RQ2, fine-tuning under the corrected, leakage-free protocol produced a cross-validated AUROC (0.661) essentially in line with the untuned validation AUROC (0.672) and the eventual test AUROC (0.675), rather than a clear improvement. The more consequential finding under this research question

was methodological rather than numerical: the initial, leakage-affected search produced a cross-validated AUROC of 0.957, an implausible figure that would have been easy to report uncritically had it not been checked against the validation-stage baseline. That the corrected figure fell back in line with the other estimates is reassuring, but the episode itself is a useful, generalizable caution about resampling techniques and cross-validation order.

On RQ3, the answer is unusually well supported for a single study, because it does not rest on this study's results alone. The test AUROC of 0.675 and Brier score of 0.095 obtained here sit almost exactly between the two independently published estimates on this same dataset, 0.667 [6] and 0.664 to 0.688 [7], despite meaningful differences in preprocessing, class-imbalance handling, and validation design across all three studies. Three independent research teams converging on the same AUROC band is stronger evidence for a genuine ceiling than any single team's robustness checks could provide on their own, and it points toward a specific, falsifiable claim: further gains on this exact prediction task are unlikely to come from a better classifier trained on the same feature set, and more likely to require additional data, laboratory trends, medication timing, social determinants, or free-text clinical notes, that this administrative dataset simply does not contain.

On RQ4, the practical answer is unambiguous. At the conventional default threshold of 0.5, the tuned model is close to useless as a screening tool, correctly identifying only 1.4% of true 30-day readmissions despite an accuracy figure that looks respectable in isolation (88.6%). Selecting a threshold via Youden's J statistic [26] on the validation set alone raises sensitivity to 58.0%, a genuinely usable operating point for a screening application, at the cost of a precision of 18.5%, meaning roughly four in five patients flagged by the model will not, in fact, be readmitted within 30 days. Whether that trade-off is acceptable depends entirely on the cost of the downstream intervention it would trigger, a judgment this study deliberately leaves to the clinical and administrative stakeholders who would deploy such a model, rather than asserting a single "correct" threshold.

5.2 Clinical and Policy Implications

Three implications follow for hospitals and health systems considering a machine learning tool of this kind for readmission risk stratification. First, model family choice matters more than model sophistication: a well-tuned gradient-boosted tree ensemble outperformed a neural network given a self-supervised pretraining advantage, and administrators should not assume that a more elaborate deep learning pipeline will outperform a properly tuned XGBoost or LightGBM model on data of this kind, particularly given the much lower computational cost of the tree-based models documented in Table 2.

Second, and more importantly, decision-threshold selection is not a cosmetic detail; it is the difference between a model that is clinically inert and one that is clinically usable, as the two rows of Table 3 make plain. Any deployment of a model like this one needs an explicit, validation-set-derived threshold chosen with the relative cost of a missed readmission against the cost of an unnecessary intervention in mind, rather than defaulting silently to 0.5, which Section 4.4 shows would be actively misleading in the opposite direction from what its accuracy figure suggests.

Third, given that three independent studies now agree on a discrimination ceiling in the neighborhood of 0.66 to 0.69 AUROC for this dataset, health systems considering a similar model for their own administrative data should calibrate expectations accordingly. This level of discrimination can meaningfully enrich a screening or triage process relative to no model at all, and the risk-stratification and feature-importance results in Section 4 remain clinically coherent, but it falls well short of the level of certainty required to make unilateral care decisions, and should be communicated to clinical staff as a triage aid rather than a diagnosis, consistent with the broader, long-standing caution in the clinical prediction modeling literature that most readmission risk is driven by factors administrative data does not fully capture [3].

5.3 Limitations

Several limitations should be kept in mind when interpreting these results. The uploaded dataset used for this study did not retain patient-level identifiers (`encounter_id` or `patient_nbr`), so the train, validation, and test split was performed at the encounter level rather than the patient level; if the same patient contributed multiple encounters, as is plausible for a chronic-disease population with an 11.4% 30-day readmission rate by definition, some encounters from the same patient could appear across different partitions, which would inflate reported performance to an unknown degree. Salim and Ibrahim [7] note the same limitation in their own use of this dataset and, notably, still arrive at closely comparable results, which is some reassurance that the effect is not large, but a patient-grouped split (via, for example, scikit-learn's `GroupKFold` [25]) would be needed to rule it out with confidence.

The self-supervised pretraining stage used a single hidden layer of 64 units and a relatively simple corruption process (additive Gaussian noise plus random masking); a deeper autoencoder, a more elaborate pretext task, such as feature-wise masked reconstruction with a learned mask rather than a fixed probability, or pretraining on a larger, related dataset before fine-tuning on this one, might have produced a more informative transferred representation and a different result for RQ1. The negative result reported here for transfer learning should therefore be read as evidence about this specific, modest implementation, not as a general claim that self-supervised pretraining cannot help tabular clinical prediction; more sophisticated tabular self-supervision methods have shown promise in other clinical prediction settings (as reviewed, for example, in recent benchmarking work on electronic health record foundation models).

This dataset spans 1999 to 2008, predating the 2012 introduction of the CMS Hospital Readmissions Reduction Program [1], which changed hospital discharge-planning incentives in ways that could shift the relationship between the predictors used here and readmission risk in more recent data; any deployment on a contemporary population would need revalidation rather than an assumption that these exact feature weights transfer forward in time. The dataset also does not distinguish planned from unplanned readmissions, and the outcome definition used here, like that used in both comparison studies in Table 4, necessarily captures both.

Finally, no subgroup analysis by race, age group, or sex was conducted here to check whether model performance is uniform across patient subpopulations. Race appeared among the top twenty predictive features in Section 4.5, which makes such an analysis a priority for any future work intending to move this pipeline toward clinical use, since differential performance across subgroups could otherwise translate a modest overall discrimination ceiling into an inequitable one. Recent fairness-focused work on related cardiometabolic readmission models illustrates why this check matters in practice: Yeh et al. [23] found that fairness-unaware readmission and mortality models for heart failure patients showed unequal false-negative rates across racial and ethnic subgroups, and that standard bias-mitigation techniques improved parity for some subgroup comparisons but not others, underscoring that subgroup performance cannot simply be assumed adequate and needs to be measured and, where necessary, corrected directly.

5.4 Future Research Directions

Several extensions follow naturally from the limitations above. Restructuring the train, validation, and test split around patient-level grouping, if patient identifiers can be recovered from the original UCI source, would close the leakage question raised in Section 5.3 directly. Enriching the feature set with data this administrative dataset does not contain, laboratory value trends across the admission, medication timing relative to discharge, or unstructured discharge-summary text processed with a clinical language model, is the most promising path past the discrimination ceiling identified in Section 4.6, since Section 5.1 and 5.2 both point toward the feature set, not the choice of classifier, as the binding constraint on further gains.

A more elaborate self-supervised pretraining scheme, deeper autoencoders, contrastive objectives, or pretraining across multiple related EHR datasets before fine-tuning on this one, would give the transfer-learning comparison in Section 4.1 a fairer test than the single-hidden-layer implementation used here, and could plausibly reverse the modest negative result reported for RQ1. Emerging tabular foundation models, pretrained transformer-style architectures that adapt to a new prediction task through in-context or retrieval-augmented inference rather than task-specific retraining, are a further direction worth testing on this exact problem; early evidence suggests such models can be competitive with gradient-boosted trees on generic tabular benchmarks but that their behavior under the real-world constraints of clinical EHR data, severe class imbalance and distribution shift among them, is not yet well characterized and would need direct evaluation here rather than being assumed [24]. Explicitly modeling the decision threshold as a cost-sensitive optimization, rather than the single Youden's-J operating point reported in Section 4.4, incorporating the institution-specific cost of a missed readmission against the cost of an unnecessary intervention, would move this pipeline from a research comparison toward a deployable clinical tool. Finally, prospective, contemporary external validation, on data collected after the 2012 policy change noted in Section 5.3, using a health system's own patient population, remains the necessary next step before any of the models compared here could be considered for real clinical deployment, consistent with the caution both comparison studies in Table 4 raise about their own results.

5. Conclusions

This study compared classical machine learning against a self-supervised, transfer-learned neural network for 30-day readmission prediction in diabetic inpatients, using the UCI Diabetes 130-US Hospitals dataset under one consistent evaluation protocol, with pretrained models, transfer learning, and data augmentation adapted to their tabular-data equivalents rather than borrowed unchanged from image or text domains. Gradient-boosted trees, XGBoost specifically, clearly outperformed both a from-scratch and a transfer-learned neural network, and a leakage-affected hyperparameter search was identified and corrected during development, a methodological episode reported here in full because the same failure mode recurs elsewhere in the class-imbalance literature. The resulting test AUROC of 0.675 sits closely alongside two independently published estimates on this identical dataset, suggesting a genuine, replicated discrimination ceiling near 0.66 to 0.69 for models built on administrative and coded clinical features alone. Most consequentially, this study shows that the conventional default decision threshold of 0.5 renders an otherwise reasonable model clinically inert, while a validation-set-derived threshold makes it usable as a screening aid at a known, quantified precision-sensitivity trade-off. The results argue for realistic, well-communicated expectations of what administrative electronic health record data alone can deliver for readmission risk stratification, and for threshold selection and self-supervised representation learning to be treated as explicit, reported methodological choices rather than defaults or afterthoughts.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analysed during the current study and analysis code are deposited on figshare at: <https://figshare.com/s/12201374caddb7bd2c40>.

Competing interests

The authors have no relevant financial or non-financial interests to disclose.

Funding

This research received no specific external funding.

Authors' contributions

A.B.: Conceptualization, Methodology, Formal analysis, Data curation, Validation, Investigation, Visualization, Writing – original draft, review & editing

B.O.: Conceptualization, Methodology, Formal analysis, Data curation, Validation, Investigation, Visualization, Writing – original draft, review & editing, Resources, Project administration.

G.O.F.: Conceptualization, Methodology, Formal analysis, Data curation, Validation, Investigation, Visualization, Writing – original draft, review & editing, Resources, Supervision, Project administration.

Acknowledgements

The authors would like to thank the Faculty of Health Studies, University of Bradford, University of Skills Training and Entrepreneurial Development (USTED), Kumasi, Ghana, and the Department of Forest Engineering, Forest Management Planning, and Terrestrial Measurements, Faculty of Silviculture and Forest Engineering, Transilvania University of Brasov, for providing some of the resources needed for this study.

References

1. Centers for Medicare & Medicaid Services. Hospital Readmissions Reduction Program (HRRP). <https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/hospital-readmissions-reduction-program-hrrp>. Accessed 10 Feb 2025.
2. Strack B, DeShazo JP, Gennings C, Olmo JL, Ventura S, Cios KJ, et al. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Res Int*. 2014;2014:781670. <https://doi.org/10.1155/2014/781670>.
3. van Walraven C, Dhalla IA, Bell C, Etchells E, Stiell IG, Zarnke K, et al. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *CMAJ*. 2010;182(6):551–7. <https://doi.org/10.1503/cmaj.091117>.
4. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32. <https://doi.org/10.1023/A:1010933404324>.
5. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2016. pp. 785 – 94. <https://doi.org/10.1145/2939672.2939785>
6. Emi-Johnson OG, Nkrumah KJ. Predicting 30-day hospital readmission in patients with diabetes using machine learning on electronic health record data. *Cureus*. 2025;17(4):e82437. <https://doi.org/10.7759/cureus.82437>.
7. Salim SS, Ibrahim AA. A machine learning approach for predicting 30-day hospital readmission in patients with diabetes. *Healthcare*. 2026;14(9):1185. <https://doi.org/10.3390/healthcare14091185>.
8. Mishra V, Tanniru MR, Sreedharan J. Prediction of 30-day readmission in diabetes management using machine learning. *Comput Biol Med*. 2025;195:110616.

<https://doi.org/10.1016/j.compbio.2025.110616>.

9. Tan J, He Y, Zhang Z, Liu J, Du J, Zhao W, et al. Derivation and validation of prediction models for prolonged length of stay and 30-day readmission in elderly patients with Type 2 diabetes mellitus: A multicenter study. *J Diabetes Res*. 2025;2025:3148242. <https://doi.org/10.1155/jdr/3148242>.
10. He Y, Yuan Y, Tan Q, Zhang X, Liu Y, Xiao M. Development and validation of a risk prediction model for 30-day readmission in elderly type 2 diabetes patients complicated with heart failure: A multicenter, retrospective study. *Front Endocrinol*. 2025;16:1534516. <https://doi.org/10.3389/fendo.2025.1534516>.
11. Wang L, Wei C, Chen Y, Lu DS, Zhang HX, Yang M. Machine learning-based prediction model for 30-day readmission risk in elderly patients with type 2 diabetes mellitus and heart failure: A retrospective cohort study with SHAP interpretability analysis. *Front Cardiovasc Med*. 2026;12:1673159. <https://doi.org/10.3389/fcvm.2025.1673159>.
12. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? *Adv Neural Inf Process Syst*. 2022;35:507–20.
13. Vincent P, Larochelle H, Bengio Y, Manzagol PA. Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th International Conference on Machine Learning*. New York: ACM; 2008. pp. 1096 – 103. <https://doi.org/10.1145/1390156.1390294>
14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic minority over-sampling technique. *J Artif Intell Res*. 2002;16:321–57. <https://doi.org/10.1613/jair.953>.
15. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146–54.
16. Zheng Y, Bensahla A, Bjelogrić M, Zaghir J, Turbe H, Bednarczyk L, et al. A scoping review of self-supervised representation learning for clinical decision making using EHR categorical data. *npj Digit Med*. 2025;8(1):362. <https://doi.org/10.1038/s41746-025-01692-1>.
17. Mataraso S, D'Souza S, Seong D, Berson E, Espinosa C, Aghaeepour N. Benchmarking of pre-training strategies for electronic health record foundation models. *JAMIA Open*. 2025;8(4):ooaf090. <https://doi.org/10.1093/jamiaopen/ooaf090>.
18. Little RJ, Rubin DB. *Statistical analysis with missing data*. 3rd ed. Hoboken: Wiley; 2019. <https://doi.org/10.1002/9781119482260>.
19. van Buuren S. *Flexible imputation of missing data*. 2nd ed. Boca Raton: CRC; 2018. <https://doi.org/10.1201/9780429492259>.
20. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res*. 2012;13:281–305.
21. Deng J, Elghobashy ME, Zang K, Patel SK, Guo E, Heybati K. So you've got a high AUC, now what? An overview of important considerations when bringing machine-learning models from computer to bedside. *Med Decis Mak*. 2025;45(6):640–53. <https://doi.org/10.1177/0272989X251343082>.
22. van Calster B, Collins GS, Vickers AJ, Wynants L, Kerr KF, Barreñada L, et al. Evaluation of performance measures in predictive artificial intelligence models to support medical decisions:

- Overview and guidance. *Lancet Digit Health*. 2025;7(12):100916.
<https://doi.org/10.1016/j.landig.2025.100916>.
23. Yeh Y, Lee Y, Huang Y, Huang W, Kimmel SE, Yang C, et al. Build fair machine learning models to predict adverse outcomes for heart failure patients with preserved ejection fraction and with reduced ejection fraction. *JAMIA Open*. 2026;9(4):ooag136.
<https://doi.org/10.1093/jamiaopen/ooag136>.
24. Pham MK, Ho TLN, Dao TTP, Mai TT, Tran MT, Ward ME et al. Retrieval-aligned tabular foundation models enable robust clinical risk prediction in electronic health records under real-world constraints. *arXiv [Preprint]*. 2026. <https://doi.org/10.48550/arXiv.2604.01841>
25. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
26. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–5.
[https://doi.org/10.1002/1097-0142\(1950\)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3](https://doi.org/10.1002/1097-0142(1950)3:1%3C32::AID-CNCR2820030106%3E3.0.CO;2-3).

Figures

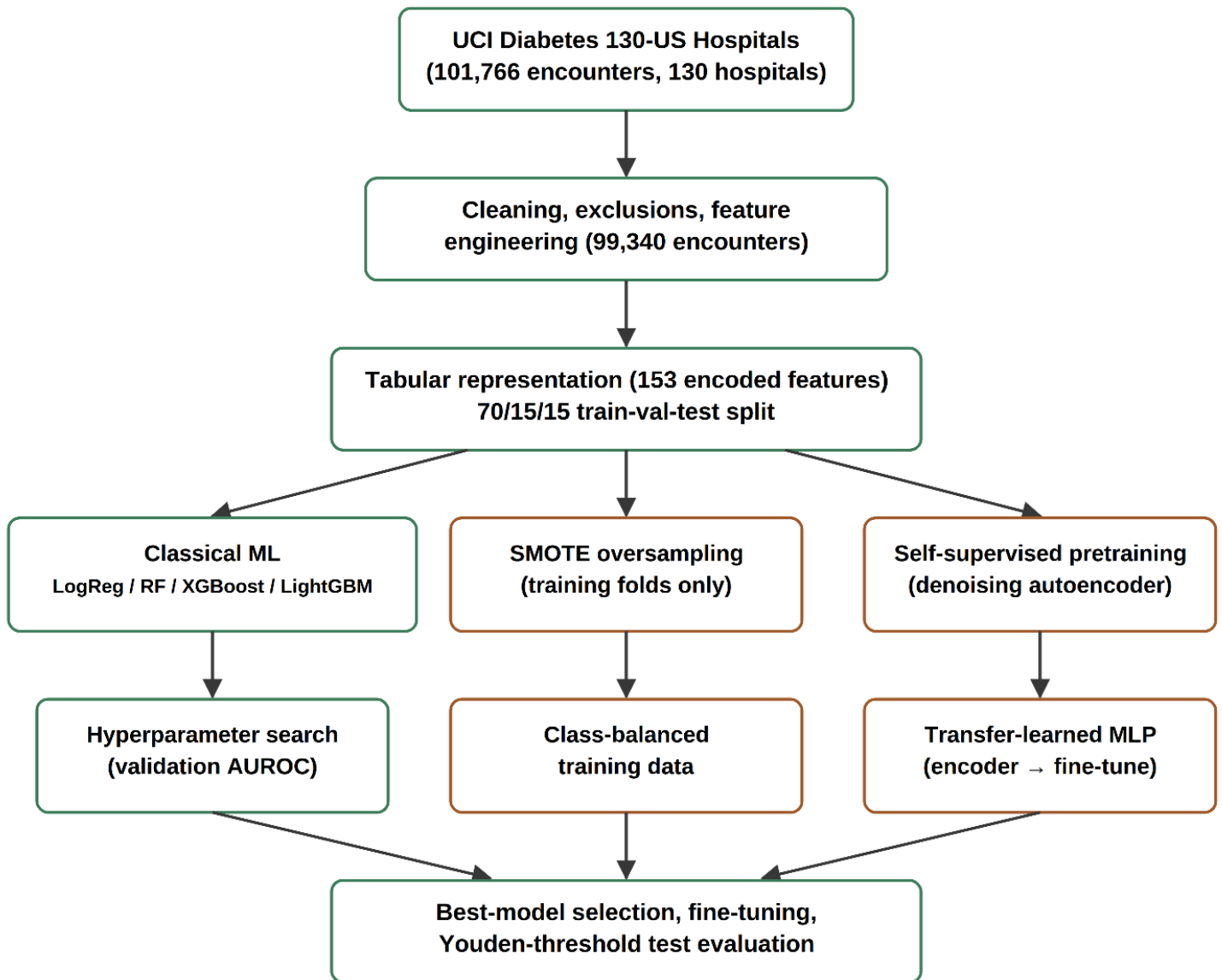


Figure 1

Conceptual framework of the readmission-prediction pipeline, from data source through tabular representation, the three model-family adaptations, and the shared evaluation and thresholding protocol.

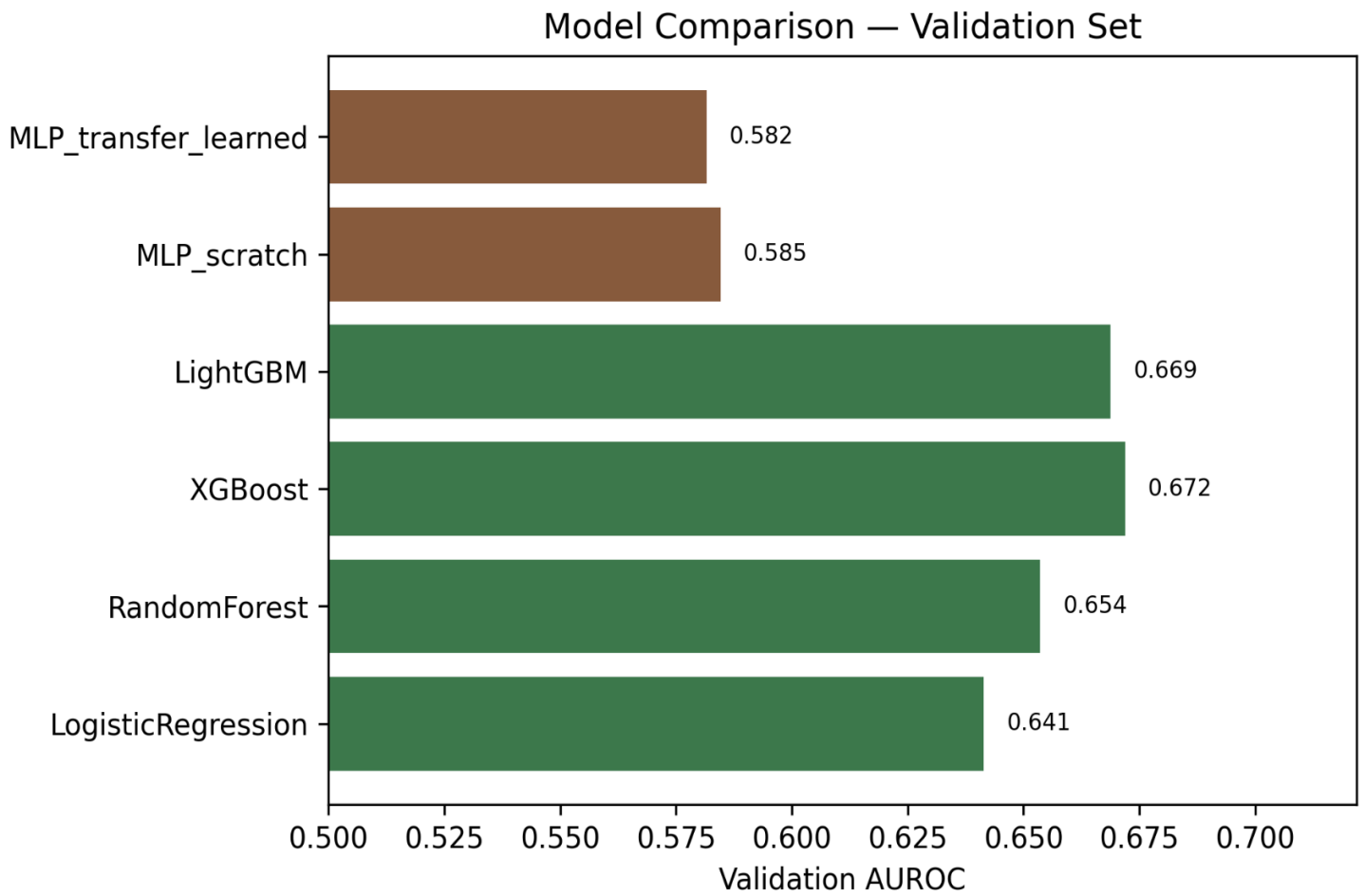


Figure 2

Validation AUROC for all six candidate models. Green bars denote tree-based and linear machine learning models; brown bars denote the two neural network variants.

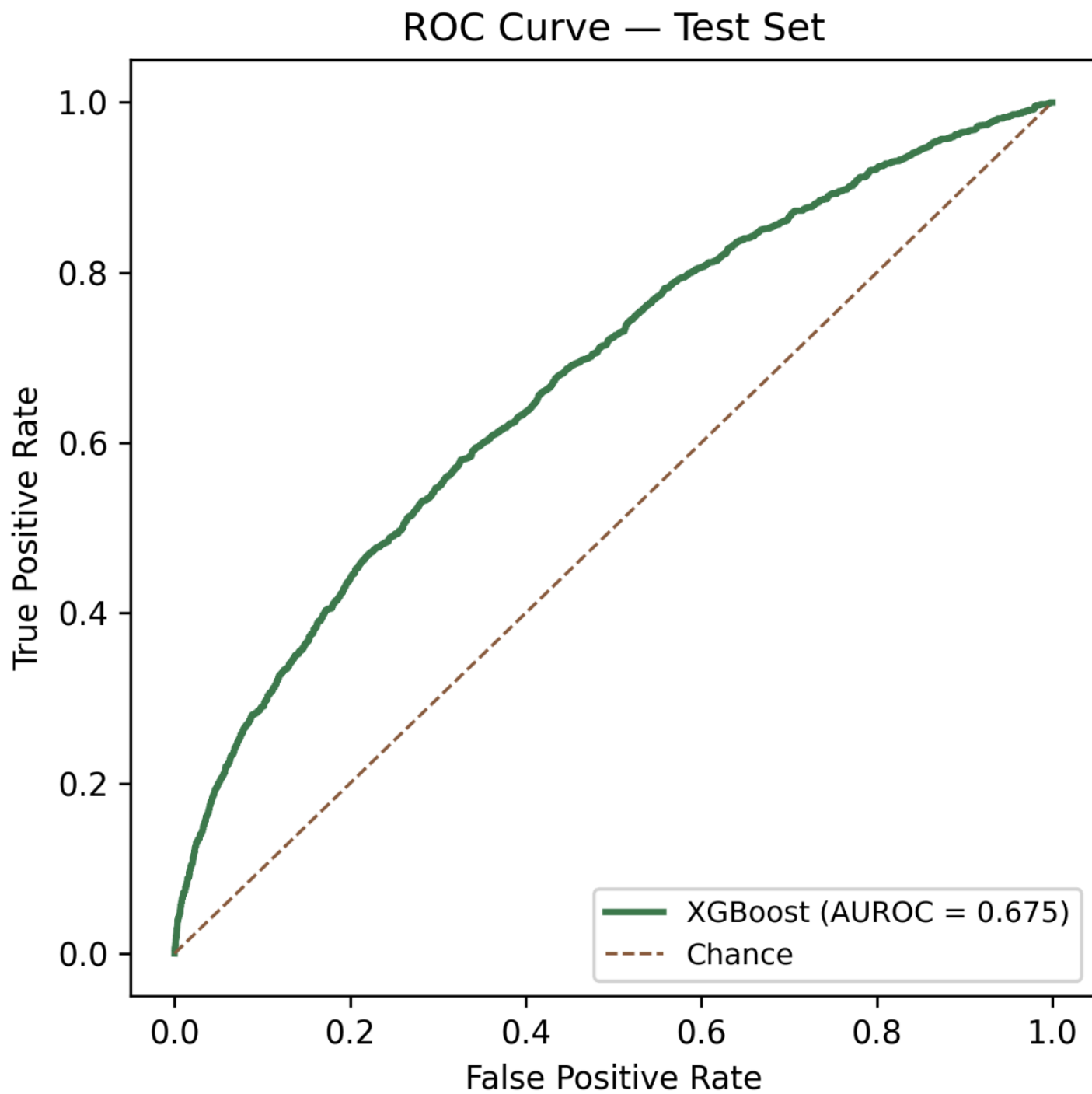


Figure 3

Receiver operating characteristic curve for the tuned XGBoost model on the test set.

Precision-Recall Curve — Test Set

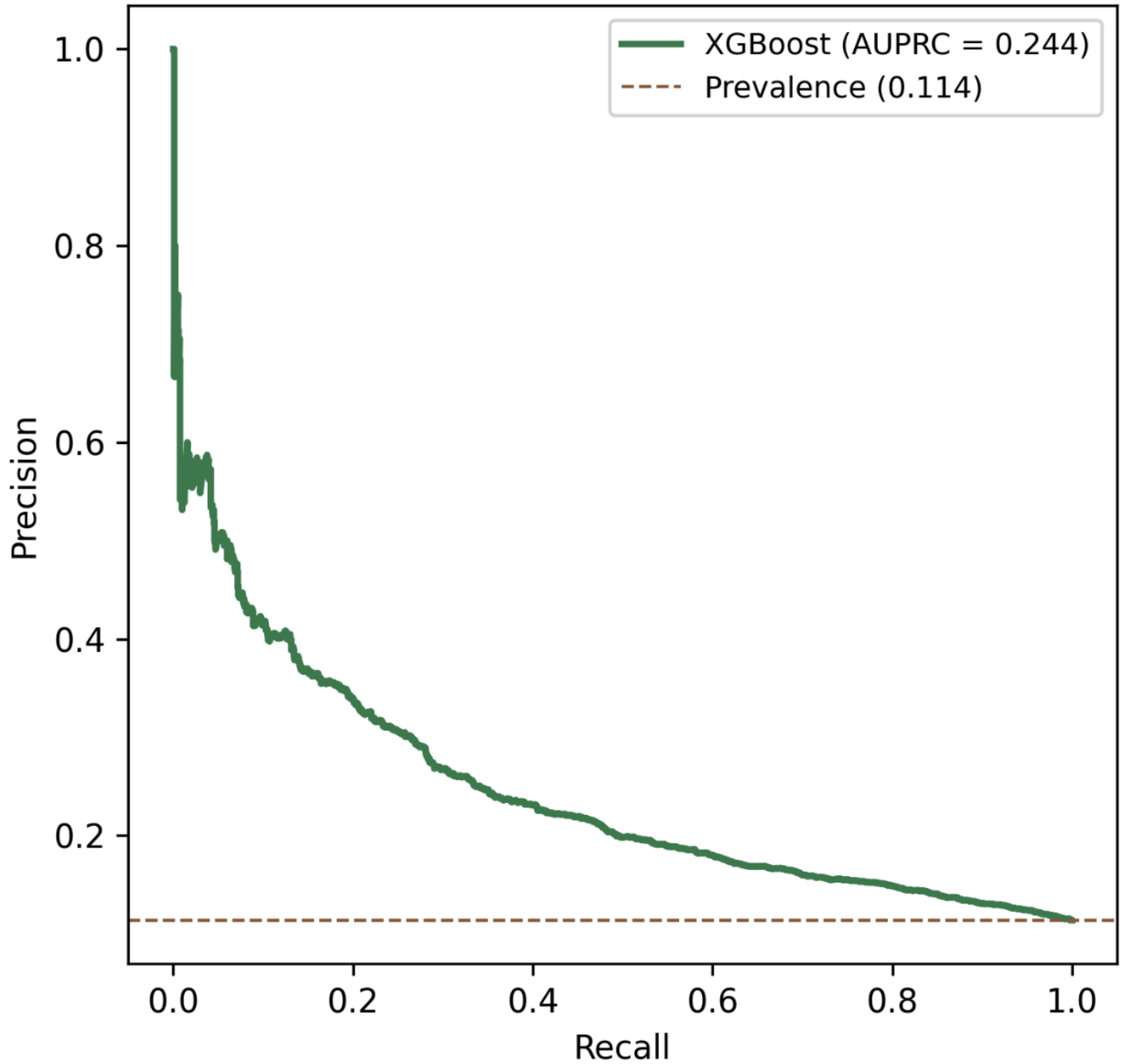


Figure 4

Precision-recall curve for the tuned XGBoost model on the test set, with the dashed line marking the baseline readmission prevalence.

Confusion Matrix (threshold=0.128)

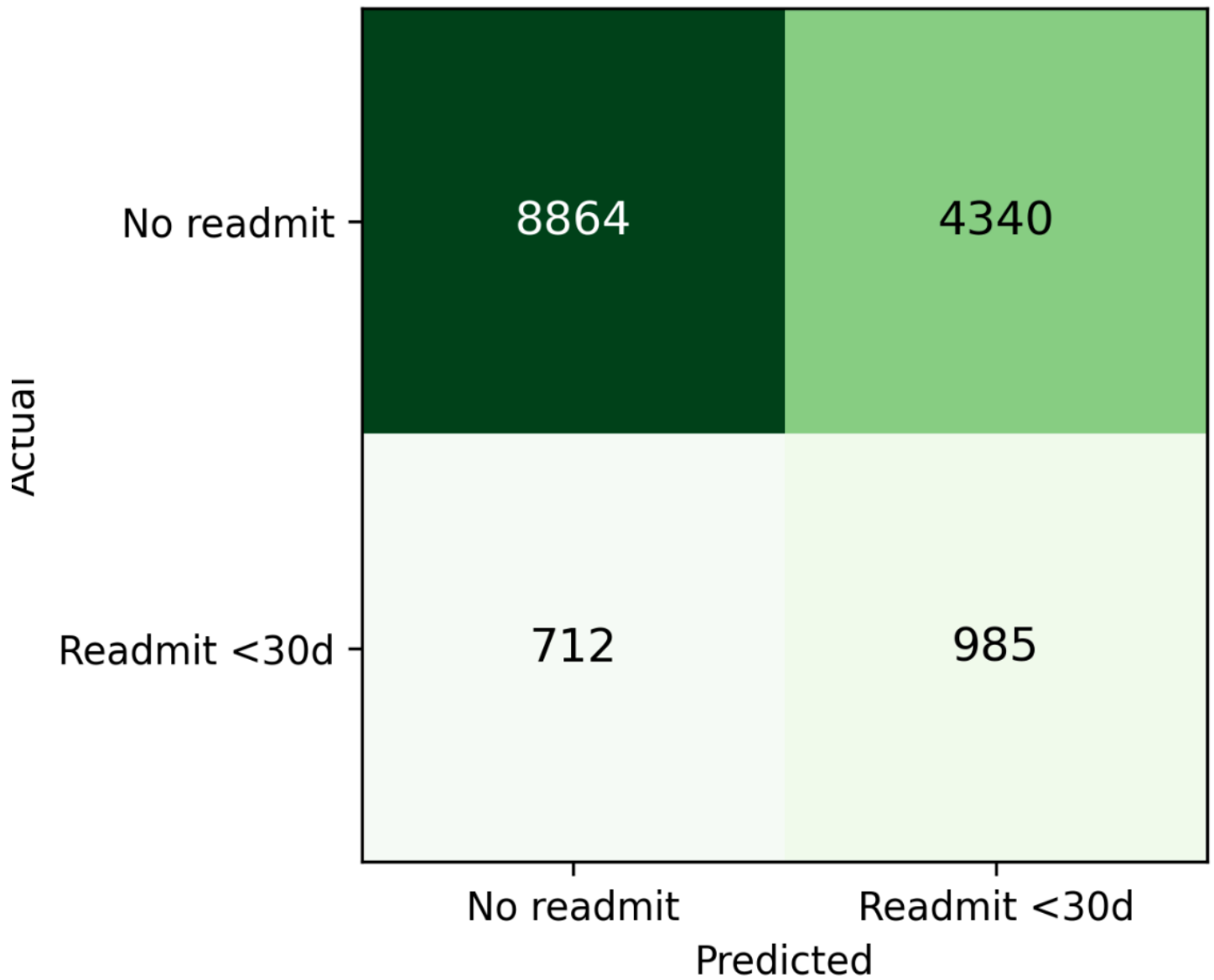


Figure 5

Confusion matrix at the Youden-optimized decision threshold (0.128), test set.

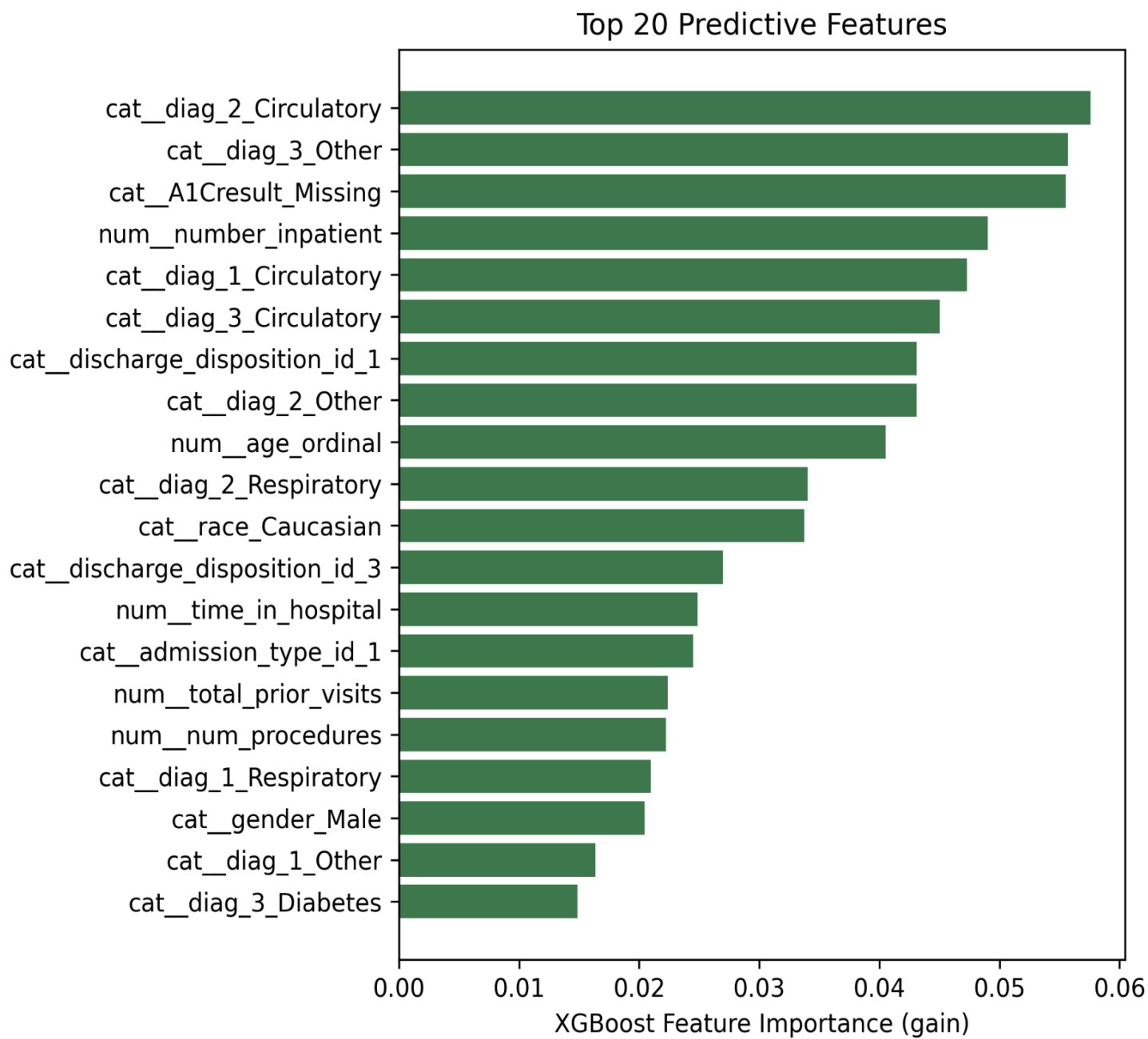


Figure 6

Top 20 predictive features by XGBoost gain, tuned model.