Predicting Hospital Readmission Risk from Electronic Health Records: A Comparative Study of Classical and Ensemble Models

Adele Chinda
Department of Computer Science
Georgia State University

Oumar Diallo
Department of Computer Science
Georgia State University

Yusuf Mumin
Department of Computer Science
Georgia State University

Abstract

We present a complete, reproducible pipeline and empirical study for predicting hospital readmission within 30 days using the *Diabetes 130-US Hospitals* dataset Clore and Strack [2014]. Our work focuses on practical preprocessing, robust imbalance handling, and a direct comparison of interpretable models (logistic regression), parametric neural models (multilayer perceptrons, MLP), and tree-ensemble approaches (XGBoost) Friedman [2001], Breiman [2001], Chen and Guestrin [2016]. The pipeline includes end-to-end engineering steps such as feature cleaning, encoding, scaling, oversampling via SMOTE Chawla et al. [2002], and feature selection using gradient-boosted gain scores. We evaluate models extensively through ROC and PR curves, loss curves, feature importance analyses, confusion matrices, and threshold tuning. We report empirical performance, highlight trade-offs between interpretability and predictive power, and discuss next steps involving temporal modeling Choi et al. [2016], Vaswani et al. [2017], calibration Zadrozny and Elkan [2001], Platt [1999], and causal evaluation Pearl [2009], Shalit et al. [2017]. The full source code and reproducibility artifacts are available at: `https://github.com/arrdel/patient-readmission-prediction`.

1 Introduction

Hospital readmission prediction is a longstanding operational and clinical challenge. Identifying patients at elevated risk allows clinicians to intervene and potentially reduce both unnecessary hospitalizations and associated healthcare costs. Electronic Health Records (EHRs) provide a rich but heterogeneous representation of patients' clinical histories, encompassing demographics, admission metadata, procedures, medications, and laboratory results. However, prediction from EHR data is complicated by missing values, noisy and high-cardinality diagnosis codes, and substantial class imbalance in short-term readmission events Hripcsak and Albers [2015].

In this study, we develop a reproducible pipeline for predicting 30-day hospital readmissions. We evaluate three modeling paradigms: logistic regression as a linear baseline, a feedforward neural network (MLP) representing nonlinear parametric methods LeCun et al. [1998], and gradient-boosted trees (XGBoost) as a strong ensemble learner for structured data Chen and Guestrin [2016]. Our contributions are: (i) a fully engineered preprocessing pipeline tailored for EHRs, (ii) a comparative experimental evaluation of classical and ensemble models, and (iii) a discussion of interpretability, operational trade-offs, and future directions.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

2 Related Work

Readmission prediction has traditionally relied on logistic regression and Cox proportional hazards models, but ensemble learners such as Random Forests Breiman [2001] and Gradient Boosting Machines Friedman [2001] have demonstrated superior predictive power in structured healthcare data. XGBoost further improves upon these methods through efficient tree boosting and regularization Chen and Guestrin [2016]. A major challenge in healthcare prediction is class imbalance; readmissions are relatively rare compared to non-readmissions. Oversampling strategies such as SMOTE Chawla et al. [2002] and its variants He et al. [2008] remain widely adopted to mitigate this issue. More advanced directions have explored neural sequence models such as RNNs and Transformers for temporal modeling of longitudinal EHRs Choi et al. [2016], Vaswani et al. [2017], as well as calibration and causal inference techniques to ensure trustworthy, actionable predictions Zadrozny and Elkan [2001], Pearl [2009].

3 Dataset and Preprocessing

We employ the publicly available *Diabetes 130-US Hospitals* dataset, consisting of 101,766 admissions across 130 hospitals from 1999–2008 Clore and Strack [2014]. The dataset comprises roughly 50 attributes, including demographics, admission types, length of stay, laboratory counts, medications, and ICD-coded diagnoses.

The prediction target is hospital readmission within 30 days. Formally, we define:

$$Y = \begin{cases} 1 & \text{if readmitted within 30 days,} \\ 0 & \text{otherwise.} \end{cases}$$

3.1 Handling Missingness

Exploratory analysis revealed high missingness in variables such as `max_glu_serum` and `A1Cresult`. Identifier fields and high-missing columns were removed, while categorical labs were imputed with explicit missing indicators. Diagnosis codes were grouped into clinically meaningful categories (e.g., circulatory, respiratory, diabetes, injury) using ICD ranges.

3.2 Feature Engineering

New features were derived to capture cumulative and rate-based behaviors. For instance, we defined the total prior visits as:

$$\text{total\_prev\_visits} = \text{number\_outpatient} + \text{number\_emergency} + \text{number\_inpatient},$$

and normalized medication usage by length of stay:

$$\text{meds\_per\_day} = \frac{\text{num\_medications}}{\text{time\_in\_hospital} + 1}.$$

Age bins were mapped to midpoint values, and binary indicators were generated for insulin use and medication changes. Numerical features were scaled with z-normalization for models requiring standardized inputs.

3.3 Feature Selection

To reduce dimensionality from one-hot encoding, we computed feature importance scores via XGBoost gain Chen and Guestrin [2016]. The top-100 features were retained for reduced experiments.

4 Methods

We compare three supervised learning paradigms that represent distinct modeling philosophies: (i) *logistic regression* as a linear and interpretable baseline, (ii) *multilayer perceptron (MLP)* as a flexible parametric neural network, and (iii) *XGBoost*, a non-parametric ensemble tree-boosting approach. This combination allows us to explore the trade-offs between interpretability, nonlinearity, and predictive performance in the context of healthcare readmission prediction.

4.1 Logistic Regression

Logistic regression (LogReg) is a generalized linear model that maps input features to a probability of readmission through a logit link function. It estimates the conditional probability of the positive class as:

$$P(Y = 1 \mid x) = \sigma(w^\top x + b),$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid activation, w denotes the coefficient vector, and b is the intercept.

The model is trained by minimizing the regularized binary cross-entropy loss:

$$\mathcal{L}_{\text{logreg}} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] + \lambda\|w\|_2^2,$$

where λ controls the strength of L2 regularization to prevent overfitting.

Logistic regression was chosen as a baseline due to its transparency: coefficients w_j can be directly interpreted as log-odds ratios, providing clinicians with an intuitive explanation of how individual risk factors contribute to readmission probability.

4.2 Multilayer Perceptron

To capture nonlinear dependencies that linear models cannot, we implemented a feedforward multilayer perceptron (MLP). The MLP applies successive affine transformations followed by nonlinear activations:

$$h_1 = \text{ReLU}(W_1 x + b_1), \quad h_2 = \text{ReLU}(W_2 h_1 + b_2), \quad \hat{y} = \sigma(W_3 h_2 + b_3).$$

Here, h_1 and h_2 are hidden representations, $\text{ReLU}(z) = \max(0, z)$ is the activation function, and the final output is squashed into $[0, 1]$ using the sigmoid.

We employed two hidden layers of size 64 and 32, balancing model expressivity with computational tractability. Training was performed using stochastic gradient descent with backpropagation and early stopping, to avoid overfitting in the presence of limited minority-class data. Unlike LogReg, MLPs are less interpretable, but they allow complex interactions between heterogeneous EHR features (e.g., nonlinear effects of age, comorbidity, and prior visits) to be modeled effectively.

4.3 XGBoost

XGBoost is a scalable and regularized gradient boosting algorithm for decision trees Chen and Guestrin [2016]. It constructs an additive model:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F},$$

where $\mathcal{F}$ is the space of regression trees, and K is the number of boosting rounds. Each tree f_k is trained to minimize the following regularized objective:

$$\mathcal{L} = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k),$$

with $\Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$ penalizing tree complexity through the number of leaves T and leaf weights w.

By sequentially fitting trees to the residuals of previous models, XGBoost effectively captures nonlinear feature interactions and hierarchical effects common in structured healthcare data. It also includes built-in handling of missing values and a class-weighting mechanism, making it well suited to clinical prediction tasks with imbalanced data.

4.4 Imbalance Handling

The dataset exhibits severe imbalance: only about 11% of admissions correspond to a 30-day readmission event. To address this, we employed the Synthetic Minority Over-sampling Technique (SMOTE)

Chawla et al. [2002], which generates synthetic minority-class samples via convex combinations of existing instances. For two minority samples x_i and x_j, a new synthetic point is created as:

$$x_{\text{new}} = x_i + \alpha(x_j - x_i), \quad \alpha \sim U(0, 1).$$

This balances the training set, allowing classifiers to learn more robust decision boundaries. In addition, for XGBoost we set the scale_pos_weight hyperparameter to:

$$\text{scale\_pos\_weight} = \frac{\#\{Y = 0\}}{\#\{Y = 1\}},$$

ensuring the loss function penalizes false negatives more heavily, thus improving sensitivity to readmission cases.

4.5 Evaluation Metrics

Evaluation of imbalanced classification requires metrics beyond raw accuracy. We report:

- ROC AUC: the area under the Receiver Operating Characteristic curve, capturing the trade-off between sensitivity and specificity.
- PR AUC: the area under the Precision–Recall curve, particularly informative when the positive class is rare.
- Precision, Recall, and F1-score for both classes, to quantify clinical trade-offs between false positives and false negatives.
- Confusion matrices and threshold-tuning plots, enabling operational analysis of classification thresholds in deployment scenarios.

This combination of threshold-independent (AUC) and threshold-dependent (precision, recall, F1) measures provides a comprehensive assessment of predictive performance in a healthcare setting.

5 Results

Figure 4 shows ROC curves comparing Logistic Regression, Multilayer Perceptron (MLP), and XGBoost on the held-out test set. Consistent with expectations for tabular EHR data, XGBoost achieved the strongest overall discriminative ability, yielding the highest ROC AUC (0.62) and PR AUC (0.42).

Table 1 summarizes the primary evaluation metrics. Logistic Regression and MLP both reached the highest F1 score (0.93), but their performance profiles differ: Logistic Regression favors higher precision (0.89) at slightly lower recall (0.97), whereas MLP achieves stronger recall (0.98) but without additional gain in precision. XGBoost, while superior in ROC/PR AUC, demonstrates a sharp trade-off, excelling in precision (0.93) but with a notably lower recall (0.40), leading to a reduced F1 score (0.56).

These differences highlight that model choice depends not only on discriminative metrics such as AUC, but also on operational trade-offs between sensitivity and specificity in deployment contexts.

Table 1: Primary evaluation metrics on the 30-day readmission task.

Model	ROC AUC	PR AUC	Precision	Recall	F1
Logistic Regression	0.56	0.35	0.89	0.97	**0.93**
MLP	0.59	0.38	0.89	**0.98**	**0.93**
XGBoost	**0.62**	**0.42**	**0.93**	0.40	0.56

6 Discussion

The results demonstrate a clear divergence between classical, neural, and ensemble approaches for hospital readmission prediction. XGBoost achieves the strongest discriminative ability, confirming

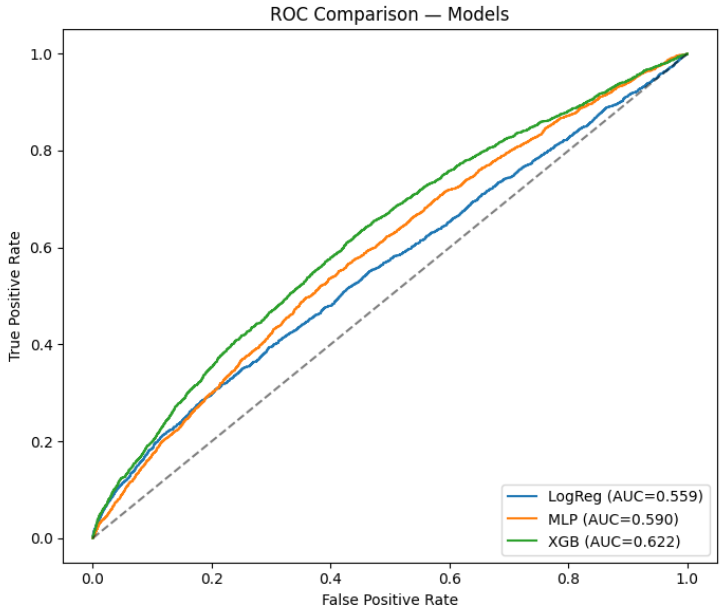

Figure 1: ROC curves comparing Logistic Regression, MLP, and XGBoost on the test set. XGBoost achieves the strongest discriminative performance (highest AUC).

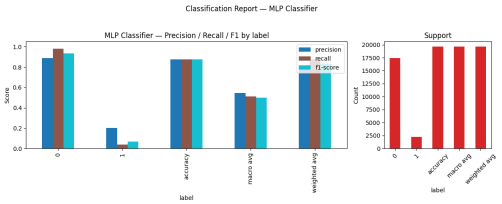

Figure 2: Classification report for MLP.

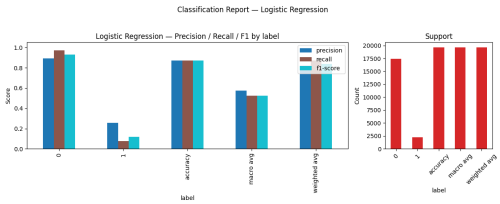

Figure 3: Classification report for Logistic Regression.

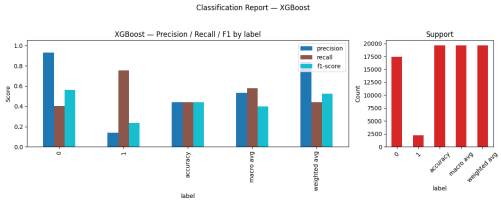

Figure 4: Classification report for XGBoost.

its status as a state-of-the-art baseline for structured medical datasets Chen and Guestrin [2016]. However, its relatively poor recall suggests that it tends to miss a large fraction of true readmissions despite being highly precise when it does predict a positive case. In clinical practice, this low sensitivity could translate into missed opportunities for intervention.

Logistic Regression and MLP, though weaker in ROC/PR AUC, exhibit a more balanced trade-off between sensitivity and precision, resulting in F1 scores (0.93) that outperform XGBoost. Logistic Regression has the added advantage of coefficient interpretability, which enables clinicians to identify and reason about feature contributions directly. MLP, while slightly more complex, demonstrates robustness with higher recall (0.98), indicating a stronger capability to capture high-risk patients, though at the cost of marginally reduced precision.

From a deployment perspective, threshold tuning becomes crucial. For instance, in high-stakes medical settings, minimizing false negatives is often prioritized, suggesting that MLP or Logistic Regression may be more suitable despite lower AUC values. Conversely, in resource-constrained contexts where unnecessary interventions are costly, the high precision of XGBoost could be preferable, provided additional measures are taken to mitigate low recall.

In summary, while XGBoost provides the best theoretical discriminative performance, Logistic Regression and MLP offer more clinically relevant trade-offs between recall and interpretability. The appropriate choice depends heavily on institutional priorities and the operational costs associated with false positives versus false negatives.

7 Limitations and Future Work

This study has several limitations. First, each admission was modeled independently, ignoring patient-level longitudinal dynamics. Future extensions should incorporate temporal models such as RNNs Choi et al. [2016] or Transformers Vaswani et al. [2017]. Second, external validation across hospitals and years remains critical before real-world deployment. Third, our models are predictive rather than causal; thus, future research should address calibration Zadrozny and Elkan [2001], Platt [1999] and causal inference frameworks Pearl [2009], Shalit et al. [2017] to evaluate intervention effects. Finally, deployment considerations such as model monitoring, data drift detection, and privacy-preserving learning are essential for practical translation.

8 Conclusions

We presented a reproducible pipeline for 30-day hospital readmission prediction. Our results highlight the strengths of ensemble methods such as XGBoost for discrimination, while also emphasizing the interpretability benefits of logistic regression. The study demonstrates how careful preprocessing, imbalance handling, and rigorous evaluation are essential to trustworthy machine learning for healthcare. Future work should integrate temporal information, calibrate predictive probabilities, and validate models externally for reliable deployment in clinical practice.

References

L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

E. Choi, A. Schuetz, W. F. Stewart, and J. Sun. Using Recurrent Neural Networks to Predict Adverse Outcomes in Intensive Care Unit Patients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2801–2809, 2016.

C. K. D. J. Clore, John and B. Strack. Diabetes 130-US Hospitals for Years 1999-2008. UCI Machine Learning Repository, 2014. DOI: https://doi.org/10.24432/C5230J.

J. H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

H. He, Y. Bai, E. A. Garcia, and S. Li. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In *IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1322–1328, 2008. doi: 10.1109/IJCNN.2008.4633969.

G. Hripcsak and D. J. Albers. Next-generation Phenotyping of Electronic Health Records. *Journal of Biomedical Informatics*, 54:195–201, 2015. doi: 10.1016/j.jbi.2015.02.002.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 9780521895606.

J. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*, pages 61–74. 1999.

U. Shalit, F. Johansson, and D. Sontag. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 3076–3085, 2017.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

B. Zadrozny and C. Elkan. Obtaining Calibrated Probability Estimates from Decision Trees and Naive Bayesian Classifiers. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 609–616, 2001.