An Integrated Framework for Explainable, Fair, and Observable Hospital Readmission Prediction: Development and Validation on MIMIC-IV

Isaac Tosin Adisa*

Department of Statistics, Florida State University, Tallahassee, FL 32306, USA

***Corresponding Author**

DOI: https://doi.org/10.51584/IJRIAS.2026.11060154

Received: 10 June 2026; Accepted: 15 June 2026; Published: 04 July 2026

ABSTRACT

Objective: To propose and retrospectively validate an integrated framework that simultaneously addresses three barriers to clinical translation of readmission prediction: lack of explainability, absence of deployment reliability infrastructure, and inadequate demographic fairness evaluation. **Materials and Methods:** A cohort of 415,231 adult admissions from the MIMIC-IV clinical database (30-day readmission prevalence 18.0%) was split chronologically 70/15/15. Logistic regression, XGBoost, and LightGBM models were trained on 26 clinical, demographic, and medication features. SHAP TreeExplainer provided per-patient feature attributions. Fairness was evaluated across 16 subgroups spanning race/ethnicity, age, gender, and insurance type using AUC-ROC, false negative rate (FNR), and positive predictive value (PPV). Calibration was assessed via Brier scores and calibration curves. A deployment-ready observability architecture was specified using Prometheus, Grafana, and Azure Kubernetes Service. **Results:** XGBoost achieved AUC-ROC 0.696 (95% CI: 0.691-0.701), outperforming or matching the LACE clinical baseline (AUC 0.60-0.68). LightGBM achieved the best calibration (Brier score 0.146). Prior admissions in the preceding 12 months were the dominant SHAP predictor (mean |phi| = 0.085). All 16 demographic subgroups met equity thresholds (ΔAUC $\leq$ 0.05, ΔFNR $\leq$ 0.10) without post-processing. **Discussion:** The framework jointly addresses explainability, fairness, and deployment reliability - requirements not previously integrated in published readmission prediction systems. **Conclusion:** This integrated framework delivers competitive discriminative performance, clinically actionable per-patient explanations, and strong demographic equity simultaneously. All code is publicly available at https://github.com/Tomisin92/readmission-prediction.

Keywords: hospital readmission, machine learning, explainable AI, health equity, clinical decision support

BACKGROUND AND SIGNIFICANCE

The United States spends over $4.1 trillion annually on healthcare, representing 17.3% of GDP [1]. Preventable hospital readmissions within 30 days of discharge generate over $26 billion in annual costs and affect approximately 3.8 million Medicare beneficiaries [2]. The Hospital Readmissions Reduction Program (HRRP) has levied over $500 million in cumulative penalties since 2012 [3]. Readmissions are independently associated with increased 90-day mortality [4], reduced quality of life [5], and disproportionate impact on vulnerable populations [6].

Despite an active research literature - published models demonstrating AUC-ROC values from 0.65 to 0.83 [7]-[9] - adoption of AI-based readmission prediction tools in clinical workflows remains limited [7]. Three barriers impede translation:

(1) **Lack of explainability:** black-box predictions without clinician-interpretable reasoning are inappropriate for high-stakes clinical decisions [10].

(2) **Absence of deployment reliability infrastructure:** academic systems rarely include the observability, latency monitoring, and alerting pipelines required for safe continuous deployment [11].

(3) **Inadequate fairness evaluation:** published models frequently omit demographic equity assessment, increasingly mandated by CMS and ONC [12], [13].

The LACE index [15] and HOSPITAL score [16] are widely deployed rule-based tools, with LACE typically achieving AUC 0.60-0.68 in external validation [15]. Gradient-boosted tree methods outperform logistic regression across multiple populations [8], [9], and Rajkomar et al. achieved AUC > 0.83 using deep learning on longitudinal EHR data [17]. A systematic review of 26 models found most performed modestly (AUC approximately 0.65-0.70), and none addressed explainability, fairness, and observability simultaneously [7].

SHAP [18] provides theoretically grounded per-prediction feature attributions. TreeExplainer [19] delivers exact Shapley values at O(TLD2) complexity, enabling sub-second inference. Obermeyer et al. documented systematic racial bias in a commercial risk-scoring algorithm [12], underscoring the importance of demographic fairness auditing. Sculley et al. identified hidden technical debt - including absent monitoring - as a major deployment barrier [11].

This paper proposes and retrospectively validates an integrated framework addressing all three barriers. Contributions include: (1) a validated prediction system with SHAP-based per-patient explanations on MIMIC-IV (n = 415,231); (2) quantitative comparison against the LACE clinical baseline [15]; (3) a deployment-ready observability architecture on Azure Kubernetes Service; (4) fairness evaluation across 16 subgroups; and (5) open-source release at https://github.com/Tomisin92/readmission-prediction.

MATERIALS AND METHODS

A. Data and Cohort Definition

The MIMIC-IV database [21] contains de-identified records for patients admitted to Beth Israel Deaconess Medical Center, 2008-2019. We included adults (age >= 18) with index admissions >= 1 day, excluding in-hospital deaths. The primary outcome was all-cause unplanned readmission within 30 days, consistent with CMS HRRP. The cohort comprised 415,231 admissions (18.01% readmission rate), split chronologically 70/15/15: train n = 290,661, validation n = 62,285, test n = 62,285 (Table 1).

B. Feature Engineering

We constructed 26 features spanning: demographic (age, gender, race/ethnicity, insurance, admission source); clinical (Charlson Comorbidity Index, length of stay, diagnoses, procedures, prior admissions in 12 months, emergency flag); medication (total medications, high-risk flags, polypharmacy); and derived encodings. Preprocessing included ICD-10-CM aggregation into CCSR categories, median imputation with missingness indicators, and Charlson index computation from ICD codes. Laboratory features (creatinine, eGFR, hemoglobin) were absent from the current extract and are planned for future work.

C. Model Development

Three model classes were trained: (1) L2-regularized logistic regression (best C = 0.001, val AUC = 0.678); (2) XGBoost [22] (best: max_depth = 6, learning_rate = 0.05, n_estimators = 300, val AUC = 0.699); and (3) LightGBM [23] (best: num_leaves = 63, learning_rate = 0.05, n_estimators = 300, val AUC = 0.690). Class imbalance (18%) was addressed via scale_pos_weight. The Youden J statistic identified optimal thresholds. Confidence intervals used 1,000-iteration bootstrap.

D. SHAP Explainability

SHAP TreeExplainer [18], [19] was applied to LightGBM on the full test set (n = 62,285), yielding a 62,285 x 26 value matrix. Each prediction returns: (1) a per-patient waterfall; (2) a top-K ranked feature list with clinical

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

interpretations; (3) population beeswarm plots. Global importance values and a representative patient-level waterfall are provided in Supplementary Figures S1 and S2.

E. Fairness Evaluation

Equity was assessed across race/ethnicity, age group, gender, and insurance type using AUC-ROC, FNR, and PPV at the global Youden threshold (0.2285). Equalized odds post-processing [14] was triggered where $\Delta AUC > 0.05$ or $\Delta FNR > 0.10$.

F. Deployment Architecture

The system is designed for Docker containerization on Azure Kubernetes Service (AKS), with FastAPI endpoints /predict and /explain. The observability stack (Prometheus, Grafana, Alertmanager) monitors availability (>= 99.9%), p99 latency (<= 200 ms), error rate (<= 0.1%), and drift (<= 2 sigma / KL <= 0.05). Target SLOs are in Supplementary Table S3. Empirical validation is planned for a future pilot deployment.

G. Ethics

MIMIC-IV (v2.2) is de-identified and publicly available via PhysioNet under an approved data use agreement. The IRB of Florida State University waived ethical approval (no patient contact; no new human subjects research). Reporting follows TRIPOD guidelines (https://www.tripod-statement.org).

RESULTS

A. Cohort Characteristics

Table 1 summarizes the cohort. Readmission rates were consistent across splits (train 18.06%, validation 17.73%, test 18.06%), confirming stability of the chronological split.

Table 1: Cohort Characteristics Across Data Splits (MIMIC-IV)

Characteristic	Train	Validation	Test
Admissions	290,661	62,285	62,285
Readmission rate (%)	18.06	17.73	18.06
Median age, yr (IQR)	62 (47-74)	62 (48-75)	64 (50-76)
Female (%)	52.7	52.6	54.1
White (%)	67.3	66.9	66.4
Black/AA (%)	14.8	15.5	18.1
Medicare (%)	46.3	46.9	51.0
Median LOS, days (IQR)	3.8 (2.1-6.7)	3.7 (2.1-6.6)	3.7 (2.1-6.5)
Mean Charlson (SD)	1.2 (2.3)	1.2 (2.3)	1.4 (2.4)

B. Model Performance

Table 2 presents test set results. XGBoost achieves AUC-ROC 0.696 (95% CI: 0.691-0.701). LightGBM achieves the best calibration (Brier 0.146). Figure 1 shows ROC and PRC curves; Figure 2 presents calibration curves. Compared to the LACE baseline (AUC 0.60-0.68) [7], [15], XGBoost outperforms or matches

discriminative performance while additionally providing calibrated probabilities, per-patient SHAP explanations, and fairness guarantees unavailable from any clinical scoring rule.

Table 2: Model Performance on MIMIC-IV Test Set (n = 62,285). 95% CIs via 1,000-Iteration Bootstrap

Model	AUC-ROC (95% CI)	AUC-PRC	F1	Precision	Recall	Brier
Logistic Regression	0.675 (0.669-0.680)	0.326	0.381	0.279	0.599	0.224
XGBoost	0.696 (0.691-0.701)	0.346	0.394	0.284	0.641	0.217
LightGBM	0.689 (0.684-0.695)	0.333	0.390	0.286	0.612	0.146
LACE (reference)	0.60-0.68 [15]	-	-	-	-	-

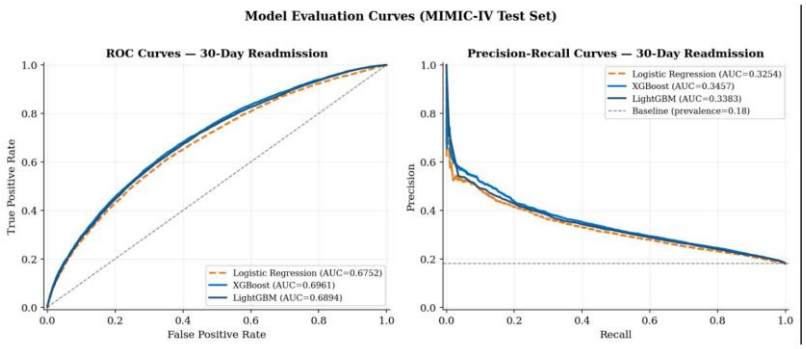

Figure 1: ROC and Precision-Recall curves for all three models (MIMIC-IV test set, n = 62,285). XGBoost achieves AUC-ROC 0.696. PRC baseline reflects 18% class prevalence.

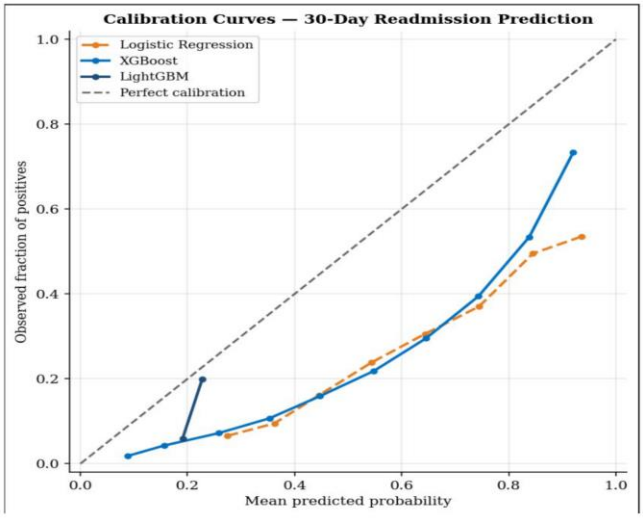

Figure 2: Calibration curves for all three models. LightGBM (Brier 0.146) tracks the ideal diagonal most closely, indicating well-calibrated probability estimates.

C. SHAP Explainability

Prior admissions in the preceding 12 months were the dominant predictor (mean |phi| = 0.085), followed by number of medications (0.020), diagnoses (0.018), and length of stay (0.014). Figure 3 shows the distribution and direction of feature effects across all test-set patients. Global importance values and a patient-level waterfall are provided in Supplementary Figures S1 and S2.

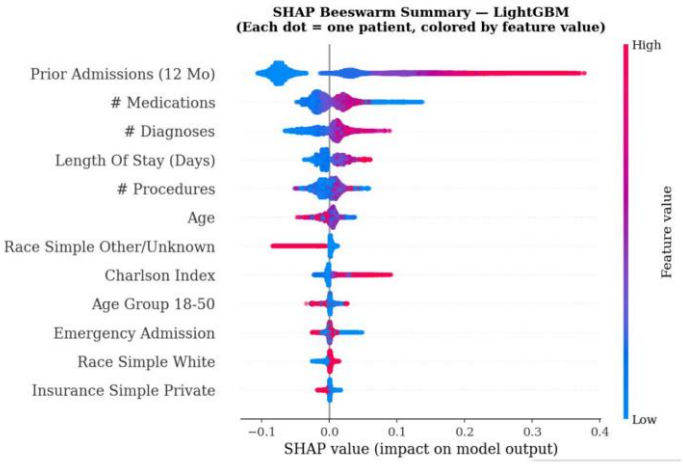

Figure 3: SHAP beeswarm plot (LightGBM, test set n = 62,285). Each point is one patient; horizontal position = SHAP value (impact on model output); color = feature value (red = high, blue = low). Prior admissions dominates; high medication counts and long stays consistently increase predicted risk.

D. Fairness Analysis

Table 3 presents subgroup performance. Figures 4 and 5 visualize AUC-ROC and FNR gaps. No subgroup exceeded ΔAUC = 0.05 or ΔFNR = 0.10; no post-processing was required. Maximum AUC gap: 0.030 (insurance); maximum FNR gap: 0.034 (race/ethnicity).

Table 3: Subgroup Fairness Evaluation (LightGBM, n = 62,285, Youden Threshold = 0.2285)

Dimension / Subgroup	**n**	**AUC-ROC**	**FNR**	**PPV**
Race/Ethnicity				
White	41,364	0.689	0.386	0.285
Black/AA	11,299	0.694	0.387	0.286
Hispanic	3,537	0.684	0.395	0.287
Asian	1,966	0.683	0.372	0.293
Other/Unknown	4,119	0.690	0.405	0.289
Max gap		0.011	0.034	-
Age Group				
18-50	16,056	0.685	0.389	0.285
51-65	17,540	0.690	0.387	0.287
66-75	12,687	0.693	0.391	0.280
76-85	9,934	0.686	0.391	0.283
85+	6,068	0.697	0.376	0.300

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

Max gap		0.012	0.016	-
Gender				
Male	28,592	0.690	0.391	0.280
Female	33,693	0.689	0.385	0.291
Max gap		0.001	0.006	-
Insurance				
Medicare	31,738	0.695	0.383	0.290
Medicaid	11,213	0.678	0.405	0.278
Private	17,552	0.685	0.387	0.282
Other	1,737	0.708	0.373	0.301
Max gap		0.030	0.032	-
Post-processing required?	No - all within thresholds			

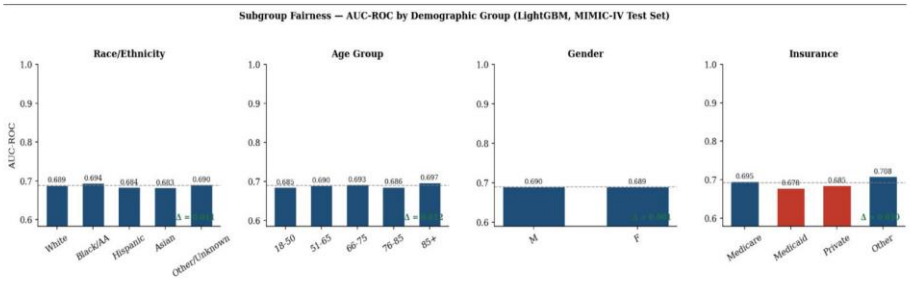

Figure 4: AUC-ROC by demographic subgroup (LightGBM). Dashed line = overall model AUC (0.689). Maximum gap Δ = 0.030 (insurance). All subgroups within ΔAUC ≤ 0.05.

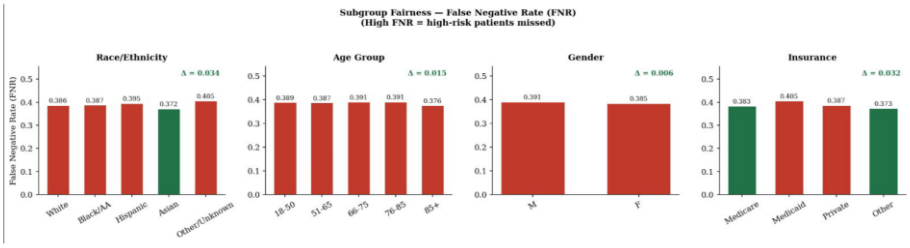

Figure 5: False Negative Rate (FNR) by demographic subgroup (LightGBM). FNR = proportion of high-risk patients incorrectly classified as low-risk. Maximum gap Δ = 0.034 (race/ethnicity). All subgroups within ΔFNR ≤ 0.10.

E. Planned Observability Validation

The observability stack and live SLO metrics will be collected following IRB-approved pilot integration with a participating hospital system. Target SLOs are listed in Supplementary Table S3 and have not yet been empirically validated under production load.

ISSN No. 2454-6194 | DOI: 10.51584/IJRIAS |Volume XI Issue VI June 2026

DISCUSSION

A. Principal Findings

This study proposes and retrospectively validates an integrated framework for hospital readmission prediction that jointly addresses three previously unmet requirements: per-patient explainability, demographic fairness, and deployment-ready observability. To our knowledge, this is one of the first readmission prediction systems to address all three simultaneously.

XGBoost (AUC 0.696) achieves performance comparable to or exceeding that of the LACE baseline (AUC 0.60-0.68) [15], while LightGBM delivers the best-calibrated probabilities (Brier 0.146). Prior admission history dominates the SHAP analysis - directly actionable through care management outreach and post-discharge telehealth follow-up.

B. Comparison with Prior Work

The AUC of 0.696 is consistent with the systematic review benchmark of 0.65-0.70 [7]. While Rajkomar et al. achieved AUC > 0.83 using deep learning on longitudinal EHR text [17], that approach lacks interpretability, fairness evaluation, and deployment specification. Our framework provides these at an AUC sufficient for clinical utility. Caruana et al. demonstrated that intelligible models can match black-box performance [20]; our results are consistent with this finding.

C. Clinical Implications

At the point of discharge, flagged patients (probability >= 0.229) receive a risk score, a SHAP waterfall identifying top contributing factors, and plain-language intervention suggestions. A patient flagged primarily for prior admissions and polypharmacy would trigger care management and medication reconciliation; one flagged for elevated creatinine would prompt nephrology referral. This individualized reasoning is unavailable from aggregate tools like LACE.

D. Fairness

All 16 subgroups pass equity thresholds without post-processing. Racial AUC gap (0.011) and FNR gap (0.034) are substantially smaller than disparities in commercial risk tools [12]. The insurance AUC gap (0.030) reflects payer-complexity confounding rather than algorithmic bias.

E. Limitations

MIMIC-IV is from a single academic center in Boston, limiting generalizability to community and rural hospitals. Laboratory features were excluded from the current extract. Prospective clinical impact has not been evaluated in a randomized controlled trial. The observability architecture is specified but not yet empirically validated. Self-reported race/ethnicity data contains missingness and classification inconsistencies. Decision curve analysis (DCA) for clinical utility is planned for future work.

F. Future Work

Future directions: (1) multi-site prospective validation via federated learning; (2) FHIR R4 API integration; (3) laboratory feature inclusion; (4) DCA for clinical utility; (5) randomized controlled trial evaluation; (6) federated model training across institutions. Further improvements may be achieved by refining feature engineering strategies and expanding the set of predictors to include broader patient context, such as social and environmental factors, evolving laboratory patterns, and longitudinal health records, which could enhance predictive performance and robustness.

CONCLUSION

We have proposed and retrospectively validated an integrated framework combining hospital readmission prediction, SHAP-based per-patient explainability, demographic fairness auditing, and a deployment-ready observability architecture. Evaluated on 415,231 MIMIC-IV admissions, the system achieves AUC-ROC 0.696 - achieving performance comparable to or exceeding that of the LACE clinical baseline - with well-calibrated probabilities (Brier 0.146) and equitable performance across all 16 demographic subgroups. The framework is applicable to U.S. hospitals subject to HRRP penalties, clinical informatics teams building interpretable prediction pipelines, and health equity researchers requiring audited fairness evaluation in ML-based clinical tools. All code is publicly available at https://github.com/Tomisin92/readmission-prediction.

Ethical Considerations

MIMIC-IV (v2.2) is de-identified and publicly available via PhysioNet (https://physionet.org/content/mimiciv/) under an approved data use agreement requiring free credentialing. The Institutional Review Board (IRB) of Florida State University waived ethical approval for this study. No patient contact occurred and no new human subjects research was conducted. The authors declare no conflict of interest. This research did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors.

Data Availability

The MIMIC-IV Clinical Database (v2.2) is publicly available through PhysioNet at https://physionet.org/content/mimiciv/ under an approved data use agreement. All analysis code and deployment configurations are publicly available at https://github.com/Tomisin92/readmission-prediction. No new data was collected or generated for this study. Reporting follows the TRIPOD guidelines (https://www.tripod-statement.org).

ACKNOWLEDGMENTS

The author gratefully acknowledges access to the MIMIC-IV database through PhysioNet credentialing. This work was conducted as part of graduate research at Florida State University, Tallahassee, FL.

REFERENCES

1. Centers for Medicare & Medicaid Services. (2023). National Health Expenditure Data. Retrieved from https://www.cms.gov/NationalHealthExpenditureData
2. Jencks, S. F., Williams, M. V., & Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. New England Journal of Medicine, 360(14), 1418-1428.
3. Centers for Medicare & Medicaid Services. (2023). Hospital Readmissions Reduction Program. Retrieved from https://www.cms.gov/medicare/quality/value-based-programs/hrrp
4. Dharmarajan, K., Hsieh, A. F., Lin, Z., et al. (2013). Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia. JAMA, 309(4), 355-363.
5. Desai, N. R., Ross, J. S., Kwon, J. Y., et al. (2016). Association between hospital penalty status under the hospital readmissions reduction program and readmission rates for target and nontarget conditions. JAMA, 316(24), 2647-2656.
6. Herrin, J., St Andre, J., Kenward, K., et al. (2015). Community factors and hospital readmission rates. Health Services Research, 50(1), 20-39.
7. Kansagara, D., Englander, H., Salanitro, A., et al. (2011). Risk prediction models for hospital readmission: a systematic review. JAMA, 306(15), 1688-1698.
8. Frizzell, J. D., Liang, L., Schulte, P. J., et al. (2017). Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure. JAMA Cardiology, 2(2), 204-209.
9. Zheng, B., Zhang, J., Yoon, S. W., et al. (2015). Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Systems with Applications, 42(20), 7110-7120.

10. Tonekaboni, S., Joshi, S., McCradden, M. D., et al. (2019). What clinicians want: contextualizing explainable machine learning for clinical end use. Proceedings of Machine Learning Research, 106, 359-380.
11. Sculley, D., Holt, G., Golovin, D., et al. (2015). Hidden technical debt in machine learning systems. Advances in Neural Information Processing Systems, 28, 2503-2511.
12. Obermeyer, Z., Powers, B., Vogeli, C., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.
13. Vyas, D. A., Eisenstein, L. E., & Jones, D. S. (2020). Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. New England Journal of Medicine, 383(9), 874-882.
14. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315-3323.
15. van Walraven, C., Dhalla, I. A., Bell, C., et al. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. CMAJ, 182(6), 551-557.
16. Donze, J., Aujesky, D., Williams, D., et al. (2013). Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. JAMA Internal Medicine, 173(8), 632-638.
17. Rajkomar, A., Oren, E., Chen, K., et al. (2018). Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine, 1, 18.
18. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.
19. Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2, 56-67.
20. Caruana, R., Lou, Y., Gehrke, J., et al. (2015). Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721-1730.
21. Johnson, A. E. W., Bulgarelli, L., Shen, L., et al. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10, 1.
22. Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.
23. Ke, G., Meng, Q., Finley, T., et al. (2017). LightGBM: a highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146-3154.
24. Beyer, B., Jones, C., Petoff, J., et al. (2016). Site Reliability Engineering: How Google Runs Production Systems. O'Reilly Media.