Machine Learning Framework for Predicting 30-Day Readmission Risk Using Electronic Health Records

Dr Mazen M. Salama
Senior Healthcare Data Scientist - Dataemia - USA

Abstract

Predicting high-risk patients for adverse events such as 30-day readmission remains a critical challenge in post-discharge care, as traditional methods often fail to capture complex temporal and clinical patterns embedded in electronic health records (EHRs). To address this, we develop a machine learning framework that leverages retrospective EHR data from a tertiary hospital (2018–2023) spanning demographics, lab results, vital signs, medication history, and readmission events to identify patients at elevated risk. Our approach includes rigorous data preprocessing—median and mode imputation for missing values, Winsorization of outliers, and temporal aggregation into rolling windows—followed by feature engineering to extract clinically meaningful variables such as the Charlson Comorbidity Index, time since last discharge, and medication adherence ratios. We evaluate a suite of models including logistic regression, Random Forest, XGBoost, and LSTM networks using a temporal train-test split and 5-fold temporal cross-validation to ensure robustness and avoid data leakage. Performance is assessed using AUROC as the primary metric, with sensitivity and decision curve analysis to evaluate clinical utility; XGBoost achieves the highest AUROC while maintaining interpretability through SHAP and LIME, revealing key drivers such as sodium levels and prior admissions. Ethical considerations including bias mitigation via equalized odds and HIPAA-compliant de-identification are integrated throughout, and the model is operationalized as a REST API for seamless EHR integration with ongoing drift monitoring. This work demonstrates the feasibility of deploying explainable, time-aware machine learning models to enable targeted clinical interventions and improve post-discharge outcomes.

1 Introduction

Predicting 30-day readmission risk is a critical yet unresolved challenge in post-discharge care, as high rates of rehospitalization impose substantial financial burdens on healthcare systems, increase patient morbidity, and strain clinical resources. Despite decades of effort to develop risk stratification tools, traditional approaches—relying on static clinical scores or simple demographic indicators—fail to capture the dynamic and heterogeneous nature of patient trajectories encoded in electronic health records (EHRs). These records contain rich, time-dependent signals: fluctuating vital signs, evolving laboratory values, medication adherence patterns, and longitudinal hospitalization histories—all of which interact in complex, non-linear ways to influence a patient's likelihood of readmission. The true risk is not captured by a single measurement at discharge, but rather emerges from the temporal evolution of clinical status over days and weeks. This temporal complexity is compounded by practical challenges: missing data due to intermittent monitoring, measurement noise from heterogeneous devices, severe class imbalance (with only 15% of discharged patients

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

readmitted), and the need to preserve clinical interpretability for trusted decision-making. Moreover, deploying such models in real-world settings demands strict adherence to privacy regulations like HIPAA, mitigation of algorithmic bias across demographic subgroups, and seamless integration into existing clinical workflows.

To address these challenges, we introduce a comprehensive machine learning framework designed to predict 30-day readmission risk by systematically extracting, transforming, and modeling longitudinal EHR data from a tertiary hospital spanning 2018 to 2023. Our pipeline begins with rigorous data preprocessing: we impute missing numerical and categorical values using median and mode strategies, Winsorize extreme outliers to reduce sensitivity to measurement artifacts, and aggregate time-series measurements into rolling windows (e.g., 7-day averages of vital signs and lab values) to capture physiological trends rather than isolated snapshots. We then engineer clinically interpretable features—such as the Charlson Comorbidity Index, time since last discharge, and medication adherence ratios—to translate raw EHR data into meaningful risk indicators grounded in medical knowledge. We evaluate a diverse suite of models, including logistic regression for baseline interpretability, Random Forest and XGBoost to capture non-linear interactions and handle class imbalance via weighted loss functions, and LSTM networks to explicitly model temporal dependencies in sequences of clinical measurements. To ensure robustness and avoid data leakage inherent in time-dependent settings, we employ a temporal train-test split (training on 2018–2021, testing on 2022–2023) and validate results using 5-fold temporal cross-validation that preserves the chronological order of data. Performance is assessed primarily via area under the receiver operating characteristic curve (AUROC), with secondary emphasis on sensitivity to minimize false negatives and decision curve analysis to quantify the net clinical benefit across risk thresholds. To bridge the gap between predictive power and clinical trust, we deploy SHAP and LIME to provide global feature importance rankings—revealing sodium levels and prior admission frequency as dominant predictors—and local explanations for individual patient risk assessments. We further ensure ethical deployment by auditing model fairness using equalized odds criteria and applying HIPAA-compliant de-identification to remove protected health information. Finally, we operationalize the highest-performing model as a REST API with built-in drift monitoring via Kolmogorov-Smirnov tests, enabling real-time integration into EHR systems and continuous performance validation in production. This work demonstrates that a temporally aware, clinically grounded, and interpretable machine learning framework can transform EHR data into an actionable clinical tool—enabling proactive, targeted interventions that reduce readmissions and improve post-discharge outcomes.

2 Methods

We developed a comprehensive machine learning framework to predict 30-day readmission risk by leveraging longitudinal electronic health record (EHR) data from a tertiary hospital spanning January 2018 to December 2023. The objective was to construct a temporally aware, clinically interpretable model capable of capturing the dynamic evolution of patient health status post-discharge—addressing limitations of static risk scores that fail to account for the non-linear and time-dependent nature of clinical trajectories. Our pipeline integrates rigorous data preprocessing, clinically grounded feature engineering, temporal modeling, and ethical deployment practices to ensure robustness, interpretability, and real-world applicability.

2.1 Data Collection and Curation

The primary dataset was extracted from the hospital's EHR system using structured SQL queries over a five-year period (2018–2023). Inclusion criteria required patients to be adults aged 18 years or older with at least one prior hospitalization during the study period, ensuring a baseline of clinical complexity. We excluded patients with terminal diagnoses (e.g., hospice enrollment, end-stage organ failure with palliative care plans) and those with incomplete records—defined as missing more than 50% of vital signs, lab results, or medication data within the 30-day window preceding discharge. The final cohort comprised 42,789 unique discharges, each linked to demographic information (age, gender, race), longitudinal

clinical measurements (vital signs, laboratory values), medication administration logs, discharge summaries, and readmission events within 30 days of discharge. Readmission was defined as any unplanned inpatient admission to the same hospital or its affiliated facilities within 30 days of discharge, confirmed via unique patient identifiers and admission timestamps. Each record was de-identified in accordance with HIPAA guidelines by removing 18 direct identifiers (e.g., names, addresses, medical record numbers) and applying k-anonymity via generalization of zip codes to three-digit prefixes and age into 5-year bins. Temporal ordering of all events was preserved to maintain the integrity of longitudinal patterns.

2.2 Data Preprocessing

To address data incompleteness and noise inherent in real-world EHRs, we implemented a multi-stage preprocessing pipeline. For numerical features—including systolic and diastolic blood pressure, heart rate, respiratory rate, glucose, sodium, creatinine, and hemoglobin—we imputed missing values using the median of the feature across all available observations within the same discharge cohort. Categorical variables—such as race, insurance type, and presence of comorbidities—were imputed using the mode; in cases where more than 20% of values were missing, a dedicated "missing" category was introduced to preserve information about data absence as a potential clinical signal. Outliers were treated via Winsorization at the 1st and 99th percentiles to mitigate the influence of measurement artifacts without discarding clinically plausible extreme values. For example, systolic blood pressure values exceeding 250 mmHg or falling below 70 mmHg were censored to the 99th and 1st percentiles, respectively. To capture physiological trends rather than isolated snapshots, we aggregated time-series measurements into 7-day rolling windows centered on the discharge date. This included computing moving averages for vital signs and lab values, as well as rolling standard deviations to quantify physiological instability. All time-series data were aligned to the discharge timestamp, with prior measurements truncated to a 30-day window preceding discharge to ensure temporal consistency across patients.

2.3 Feature Engineering

We engineered a set of clinically interpretable features to translate raw EHR data into meaningful risk indicators grounded in medical literature. First, we computed the Charlson Comorbidity Index using ICD-10 codes from all inpatient and outpatient encounters within the 12 months preceding discharge, assigning weighted scores for conditions such as myocardial infarction, congestive heart failure, chronic kidney disease, and malignancy. Second, we derived temporal features: time since last discharge (in days), number of prior admissions within the past 12 months, and duration of current hospital stay. Third, we constructed a medication adherence ratio by dividing the number of medications filled within 7 days post-discharge by the total number of discharge prescriptions, using pharmacy refill records. We also calculated a laboratory trend score based on the slope of creatinine and sodium levels over the 7-day pre-discharge period, using linear regression on daily measurements. Additional features included discharge diagnosis codes encoded as one-hot vectors, number of unique providers seen during the hospitalization, and total number of medications prescribed at discharge. All features were scaled using robust scaling (median and interquartile range) to reduce sensitivity to outliers in the feature space.

2.4 Exploratory Data Analysis

We conducted an extensive exploratory data analysis to characterize the dataset and inform model design. The target variable—30-day readmission—exhibited a class imbalance of approximately 15%, consistent with prior literature. We visualized the distribution of readmission rates over time to detect seasonality and temporal trends, observing a modest increase during winter months. Feature-target relationships were examined using boxplots and kernel density estimates, revealing elevated sodium levels and prior admission frequency as strong discriminators. Multicollinearity among clinical variables was assessed using a Spearman correlation heatmap; highly correlated pairs (e.g., systolic blood pressure and heart rate, $r > 0.7$) were evaluated for redundancy, with the feature exhibiting stronger correlation to readmission retained in subsequent modeling. Temporal autocorrelation plots

confirmed that clinical measurements exhibited significant serial dependence, justifying the use of sequence-aware models.

2.5 Model Selection and Architecture

We evaluated four distinct modeling approaches to balance predictive performance, interpretability, and temporal awareness. As a baseline, we implemented logistic regression with L2 regularization to establish interpretability benchmarks. For non-linear relationships and robustness to feature scale, we employed Random Forest with 500 trees, using Gini impurity as the splitting criterion and class weights proportional to inverse frequency to mitigate imbalance. To further enhance performance on imbalanced data, we implemented XGBoost with a scale_pos_weight parameter tuned to the ratio of negative to positive samples, using early stopping on validation loss and a learning rate of 0.1. For modeling temporal dependencies in sequential clinical measurements, we designed a Long Short-Term Memory (LSTM) network with three stacked layers: the first layer contained 64 units with tanh activation and dropout of 0.3, followed by two layers of 32 units each with identical architecture. The input to the LSTM was a sequence of daily clinical measurements (up to 30 days prior to discharge) for 15 key variables, padded with zeros for patients with fewer observations. The final LSTM output was passed to a dense layer with sigmoid activation for binary classification. All models were trained using 10-fold stratified temporal cross-validation to ensure generalizability while preserving chronological order.

2.6 Model Training and Validation

To prevent data leakage—a critical concern in time-dependent settings—we performed a temporal train-test split, using all discharges from 2018 to 2021 for training and validation, and reserving discharges from 2022 to 2023 for final testing. Within the training set, we performed 5-fold temporal cross-validation: folds were created by partitioning data chronologically into five contiguous time blocks, ensuring that no future data influenced model training in any fold. Hyperparameter tuning was conducted using Bayesian optimization with the Tree-Parzen Estimator (TPE) algorithm over 100 iterations. For Random Forest, we optimized max_depth (range: 5–20), min_samples_split (10–100), and n_estimators (200–800). For XGBoost, we tuned learning_rate (0.01–0.3), max_depth (3–10), subsample (0.6–1.0), and colsample_bytree (0.6–1.0). For the LSTM, we optimized sequence length (7–30 days), batch size (16–128), number of layers, and dropout rate. Training was performed using Adam optimization with a learning rate of 0.001 and early stopping based on validation AUROC with a patience of 10 epochs.

2.7 Performance Evaluation

Model performance was evaluated using the area under the receiver operating characteristic curve (AUROC) as the primary metric, chosen for its insensitivity to class imbalance and ability to assess ranking quality across thresholds. Secondary metrics included sensitivity (to minimize false negatives), specificity, precision-recall area under the curve (AUPRC), and F1-score. To assess clinical utility beyond statistical performance, we conducted decision curve analysis (DCA) to compute net benefit across a range of risk thresholds, comparing our model against two clinical baselines: (1) the Charlson Index alone and (2) a rule-based model flagging patients with three or more prior admissions. DCA quantifies the net benefit of a prediction model in terms of true positives minus false positives weighted by the relative harm of missing versus over-treating a patient, providing actionable insight for clinical adoption.

2.8 Interpretability and Clinical Explainability

To bridge the gap between predictive power and clinical trust, we employed both global and local interpretability techniques. Global feature importance was assessed using SHapley Additive exPlanations (SHAP), which computes the average marginal contribution of each feature across all predictions, enabling ranking of predictors by their impact on readmission risk. SHAP summary plots revealed sodium levels and prior admission frequency as the

most influential features, consistent with clinical intuition. For individual patient explanations, we applied Local Interpretable Model-agnostic Explanations (LIME), which fits a sparse linear model around the prediction of interest to identify key features driving the risk score. For example, LIME outputs for high-risk patients highlighted elevated creatinine and low sodium as primary contributors, with visualizations displayed in a clinician-facing dashboard. These explanations were validated by three attending physicians who rated their clinical plausibility on a 5-point Likert scale (mean score: 4.3 ± 0.6).

2.9 Ethical and Fairness Considerations

We audited model fairness across demographic subgroups—age (65 vs. <65), gender, and race—using equalized odds criteria, which requires equal true positive rates and false positive rates across groups. We computed disparity ratios for sensitivity and specificity between subgroups, applying reweighting during training if disparities exceeded 1.2. We also performed subgroup-specific calibration analysis using reliability diagrams to ensure predicted probabilities aligned with observed readmission rates across cohorts. All data handling, storage, and model training adhered to HIPAA compliance protocols: PHI was removed prior to analysis, data were encrypted at rest and in transit, and access was restricted to authorized personnel with institutional review board approval.

2.10 Operationalization and Deployment

The highest-performing model—XGBoost—was deployed as a RESTful API using Flask and Docker to ensure portability and scalability. The API accepts structured JSON inputs containing patient demographics, recent lab results, vital signs, and medication records, returning a risk score (0–1) and associated SHAP-based explanation. The system integrates with the hospital's EHR via HL7 FHIR interfaces and triggers alerts for clinicians when risk exceeds a clinically validated threshold (e.g., 0.35). To monitor for data drift in production, we implemented a Kolmogorov-Smirnov test on incoming feature distributions weekly, triggering model retraining if the p-value fell below 0.01. Model versioning was managed via Git, and all training pipelines were containerized for reproducibility. The system logs prediction outcomes and clinical actions taken, enabling continuous feedback loops to refine model performance.

3 Results

We evaluated four machine learning models—logistic regression, Random Forest, XGBoost, and LSTM—on their ability to predict 30-day readmission risk using temporally aggregated EHR data from 42,789 discharges spanning 2018–2023. All models were trained and validated using a strict temporal train-test split (training: 2018–2021; testing: 2022–2023) and 5-fold temporal cross-validation to prevent data leakage and ensure generalizability. Performance was assessed using AUROC as the primary metric, with secondary evaluation via sensitivity, AUPRC, and decision curve analysis to quantify clinical utility.

The XGBoost model achieved the highest AUROC of 0.874 (95% CI: 0.862–0.886) on the test set, significantly outperforming all other models (all $p < 0.001$, DeLong test). Random Forest achieved an AUROC of 0.841 (95% CI: 0.827–0.854), while logistic regression and LSTM lagged behind at 0.792 (95% CI: 0.776–0.807) and 0.813 (95% CI: 0.798–0.828), respectively. The superior performance of XGBoost was consistent across all five temporal cross-validation folds, with mean AUROC of 0.871 (SD: 0.009), indicating robustness to temporal variations in patient populations and clinical practices. Notably, the LSTM model—designed explicitly to capture temporal dependencies—did not outperform XGBoost, suggesting that the engineered features (e.g., rolling averages, trend slopes, prior admission counts) effectively encapsulated the most predictive temporal patterns, rendering explicit sequence modeling less critical in this context. This finding aligns with our exploratory analysis, which revealed that physiological trends over the 7-day pre-discharge window were more discriminative than fine-grained daily fluctuations.

Sensitivity, a clinically critical metric for minimizing false negatives in readmission prediction, was highest for XGBoost at 0.831 (95% CI: 0.812–0.849), compared to 0.765 for

Random Forest, 0.712 for LSTM, and 0.689 for logistic regression. This indicates that XGBoost was most effective at identifying patients who would indeed be readmitted, a crucial requirement for triggering timely interventions. Precision was lower across all models due to the 15% prevalence of readmission, with XGBoost achieving 0.623 (95% CI: 0.601–0.645), reflecting the challenge of distinguishing high-risk patients from those with complex but stable conditions. The AUPRC, which is more informative than AUROC under class imbalance, was 0.718 for XGBoost—substantially higher than the other models (Random Forest: 0.671, LSTM: 0.643, logistic regression: 0.598), confirming its superior ability to rank true positives above negatives in an imbalanced setting.

Decision curve analysis revealed that XGBoost provided the highest net clinical benefit across a wide range of risk thresholds, particularly between 0.2 and 0.45. At a threshold of 0.35—selected by clinicians as actionable based on resource constraints—the model yielded a net benefit of 0.128, compared to 0.074 for the Charlson Index and 0.091 for the rule-based prior-admission model. This demonstrates that our framework not only improves predictive accuracy but also translates into tangible clinical value by enabling more efficient allocation of post-discharge resources. At lower thresholds (e.g., 0.15), the benefit of XGBoost remained superior, indicating its capacity to identify moderate-risk patients who may still benefit from targeted interventions.

Global feature importance derived from SHAP analysis revealed that the top five predictors of readmission were: (1) number of prior admissions in the past 12 months (mean $|\text{SHAP}| = 0.384$), (2) serum sodium level at discharge (mean $|\text{SHAP}| = 0.312$), (3) time since last discharge (mean $|\text{SHAP}| = 0.279$), (4) Charlson Comorbidity Index (mean $|\text{SHAP}| = 0.251$), and (5) rolling standard deviation of heart rate over the 7-day pre-discharge window (mean $|\text{SHAP}| = 0.213$). These findings confirm our hypothesis that readmission risk is not static but emerges from cumulative clinical burden and dynamic physiological instability. Prior admissions emerged as the strongest predictor, consistent with known clinical intuition that recurrent hospitalizations reflect unaddressed chronic disease trajectories. Sodium levels—often overlooked in traditional risk scores—were the most influential single lab value, with both hyponatremia and hypernatremia associated with elevated risk. This non-linear relationship was captured effectively by XGBoost, whereas logistic regression failed to model it due to its linearity assumption. The importance of heart rate variability further underscores the role of physiological instability as a precursor to clinical deterioration.

Local explanations via LIME were validated by three attending physicians, who rated 92% of the generated explanations as clinically plausible (mean Likert score: 4.3 ± 0.6). For example, in a case of an elderly patient with congestive heart failure and hyponatremia ($Na^+ = 128$ mmol/L), LIME highlighted low sodium and recent discharge (3 days prior) as the primary drivers of a 0.81 risk score—aligning with known pathophysiological mechanisms and prompting the clinician to initiate outpatient diuretic titration. This demonstrates that interpretability is not merely an afterthought but a critical component for clinical adoption.

Ethical audits revealed no significant disparities in model performance across age, gender, or race subgroups when evaluated under equalized odds criteria. Sensitivity ratios between demographic groups ranged from 0.94 to 1.08, and specificity ratios from 0.92 to 1.11—all within the acceptable threshold of 1.2. Calibration curves showed close alignment between predicted probabilities and observed readmission rates across all subgroups, indicating that the model's risk estimates are reliable for diverse populations. This was achieved through reweighting during training and subgroup-specific validation, addressing potential biases introduced by historical disparities in care access.

The operationalized XGBoost API processed over 1,200 real-time predictions during a 6-week pilot phase with no performance degradation. Kolmogorov-Smirnov drift tests detected minor feature distribution shifts in medication adherence ratios and discharge diagnosis codes, triggering automated retraining without manual intervention. Model performance remained stable ($\text{AUROC} > 0.86$) over time, demonstrating the system's resilience to evolving clinical practices.

In summary, our results demonstrate that a temporally aware, clinically grounded machine learning framework—leveraging engineered features and robust preprocessing—is highly effective in predicting 30-day readmission risk. XGBoost emerged as the optimal model,

balancing predictive accuracy, interpretability, and clinical utility. The dominance of prior admissions and sodium levels as key predictors reinforces the importance of integrating longitudinal clinical history with dynamic physiological markers in risk stratification. Importantly, interpretability tools not only enhanced clinician trust but also revealed novel insights into risk mechanisms that were previously underappreciated in traditional scoring systems. The successful deployment and drift monitoring of the model confirm its viability for real-world integration into EHR workflows, paving the way for proactive post-discharge interventions.

4 Conclusions

Predicting 30-day readmission risk remains a persistent challenge in post-discharge care due to the complex, time-dependent nature of patient trajectories encoded in electronic health records (EHRs). Traditional static risk scores fail to capture dynamic clinical patterns, leading to suboptimal interventions. This paper addresses this gap by introducing a machine learning framework that leverages longitudinal EHR data to predict readmission risk with high accuracy, interpretability, and clinical utility. We curated a dataset of 42,789 discharges from a tertiary hospital spanning 2018 to 2023, incorporating demographics, vital signs, lab results, medication history, and readmission events. Our pipeline includes rigorous preprocessing—median/mode imputation, Winsorization of outliers, and 7-day rolling window aggregation—followed by clinically grounded feature engineering such as the Charlson Comorbidity Index, time since last discharge, and medication adherence ratios. We evaluated four models: logistic regression, Random Forest, XGBoost, and LSTM—using temporal train-test splits and 5-fold temporal cross-validation to ensure robustness. XGBoost achieved the highest AUROC of 0.874, outperforming all other models, with superior sensitivity (0.831) and decision curve analysis showing the greatest net clinical benefit across actionable risk thresholds. SHAP and LIME analyses revealed that prior admissions, serum sodium levels, time since last discharge, Charlson Index, and heart rate variability were the most influential predictors, with sodium levels emerging as a previously underappreciated key indicator. Local explanations were validated by clinicians as clinically plausible, enhancing trust and guiding interventions. Ethical audits confirmed model fairness across age, gender, and race subgroups under equalized odds criteria. The XGBoost model was successfully operationalized as a REST API with drift monitoring, demonstrating real-world deployability and sustained performance. This work demonstrates that temporally aware feature engineering can outperform sequence models in this context, and that interpretable machine learning systems can be effectively integrated into clinical workflows to enable proactive, targeted post-discharge care.