Deployment-Aware 30-Day Readmission Prediction in Resource-Limited Hospitals: Calibration, Threshold Policy, and Decision Utility

Samer Asad Malalha[1], Ma Burhanuddin[2*], Hatem T M Duhair[3], Jamil Abedalrahim Jamil Alsayaydeh[4], Mazen Farid[5]
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka (UTeM), 76100 Melaka, Malaysia[1, 2]
Department of Engineering Technology, Fakulti Teknologi Dan Kejuruteraan Elektronik Dan Komputer (FTKEK), Universiti Teknikal Malaysia Melaka (UTeM), 76100 Melaka, Malaysia[3, 4]
Faculty of Information Science and Technology (FIST), Multimedia University, Melaka 75450, Malaysia[5]
Centre for Intelligent Cloud Computing, COE for Advanced Cloud, Multimedia University, Melaka 75450, Malaysia[5]

***Abstract*—Thirty-day hospital readmission is a well-established quality metric, and many clinical prediction models have been developed for this task; however, high discrimination does not by itself mean that a model is safe to use in discharge workflows. This study developed and applied an integrated deployment-oriented evaluation workflow in which calibration, threshold governance, and decision utility were treated as primary evaluation requirements rather than secondary diagnostics. Retrospective inpatient data collected between 2022 and 2024 from two resource-limited government hospitals were used (N = 30,000; readmission prevalence = 15.0%). The analysis was based on patient-level internal validation using non-overlapping training, validation, and held-out test partitions. A multilayer perceptron neural network and a random forest were evaluated using patient-level grouping (70% training, 15% validation, 15% test). Both models showed strong discrimination on the held-out test set (ROC-AUC = 0.868 for the neural network and 0.880 for the random forest), with nearly identical minority class detection (PR-AUC = 0.461 vs 0.460). However, calibration analyses separated the models despite identical Brier scores (0.095). The neural network showed lower expected calibration error (ECE = 0.013 vs 0.035) and near ideal probability scaling (calibration slope = 0.971, 95% CI: [0.94, 1.00]) compared with the random forest (slope = 1.202). Threshold analysis also showed that a default threshold could be unsafe, since recall was 0.244 at 0.50 but increased to 0.867 at 0.12, while false negatives dropped from 510 to 90. Decision Curve Analysis further supported the neural network, including a mean net benefit of 0.097 at a threshold of 0.12. Practically, threshold, model version, and monthly calibration summaries should be logged in an audit trail.**

Keywords*—Clinical AI deployment; hospital readmission prediction; expected calibration error; electronic health records; operating region; probability reliability; model governance

I. Introduction

The use of thirty-day hospital readmission as a measure of quality of care and as a target for predictive analytics in numerous health systems has been well-documented [1]-[4]. The increasing availability of electronic health records has allowed machine learning and deep learning models to process large amounts of high-dimensional clinical information [5], [6]. However, the field still faces a similar translation problem: models that perform well retrospectively do not always act safely or consistently when embedded in the workflow of discharge planning, care coordination, and follow-up allocation [7], [8].

Discrimination metrics, specifically the area under the receiver operating characteristic (ROC) curve, dominate model evaluation in the readmission prediction literature [9], [10]. While discrimination is valuable because it measures a model's ability to rank patients based on their risk, it does not necessarily imply that the model produces numerically valid predicted probabilities that can inform decisions regarding absolute risk [11], [12]. Therefore, methodological critiques have emphasized that clinical risk prediction needs to evaluate model performance using additional methodologies that assess the model's ability to produce valid predicted probabilities and that examine the decision-making consequences of model predictions in the context of realistic operating conditions [3], [12] – [15].

Calibration is central to this deployment question. Calibration occurs when the model predicts the actual rates of events; therefore, the predicted risk is expressed as a frequency statement in the population of interest [12], [13]. If calibration fails, then risk will be either systematically over- or under-predicted, and downstream decisions will potentially result in under-treatment, over-treatment, or misallocation of resources, regardless of whether the discrimination level of the model remains unchanged. These problems are not theoretical. Miscalibration can affect what patients receive: intensive discharge counseling, early outpatient follow-up, or post-discharge monitoring. As previously described, these interventions can directly affect patient safety and workload in environments where the margin is thin [8], [16]. Calibration monitoring has recently been added to the list of lifecycle management responsibilities through recent governance efforts, and is no longer viewed as a one-time diagnostic during model development [17], [18].

There exists a second type of deployment constraint related to converting predicted risk into actionable decisions. The current practice of setting a default threshold of .50 for binary classification in order to classify a patient as being at high risk of readmission is increasingly difficult to defend in clinical settings, since the choice of threshold rarely corresponds to clinical preference, the operational capacity of the institution, or

*Corresponding author.

the relative cost of different types of errors [7], [19]. Choice of threshold determines the balance between false negatives and false positives, and this balance has significant clinical implications for preventing readmissions. For example, missing a high-risk patient may prevent the triggering of a follow-up pathway that could have prevented the patient's deterioration, while flagging too many low-risk patients may lead to saturation of available staff capacity and reduce the focus on those most likely to benefit from follow-up. An operational way to address this issue is to document a threshold selection process based upon institutional capacity, e.g., select a threshold that would yield approximately 25 to 30 flagged discharges per week because that is the number of transitional care calls that a team can reliably make. During deployment, the selected threshold should be recorded along with the model version and the effective date so that changes in decision-making behavior can be tracked over time. This framing of threshold optimization as a clinical design decision instead of a statistical convention provides an opportunity to link clinical design to the technical capabilities of the model [20]-[22].

Decision Curve Analysis (DCA) provides a mechanism for evaluating whether action on a model improves outcome, across a range of clinically relevant thresholds, by explicitly incorporating the relative harms of false positives and false negatives into net benefit [23], [24]. Although DCA is increasingly being advocated for in guidelines for the evaluation and validation of prediction models [14], [25], it is relatively rare in the evaluation of hospital readmission prediction models, where evaluation typically focuses on discrimination and limited calibration metrics [3], [8]. This gap is important because a model can appear to be highly discriminative on ROC-AUC, yet fail to provide decision utility in the threshold ranges that are relevant to discharge workflows.

These limitations are accentuated in resource-constrained systems where staffing shortages, fragmented documentation, and limited follow-up capacity can amplify the effects of predictive error [2], [7]. In such settings, a deployment-ready readmission model should exhibit more than just discrimination. It should demonstrate stable calibration and predictable threshold behavior, so that the workflow consequences of acting on the model remain understandable and manageable.

Therefore, motivated by the need to evaluate readmission models in terms that are closer to clinical use, the current study develops and applies an integrated deployment-oriented evaluation workflow that combines discrimination, calibration reliability, threshold sensitivity analysis, and decision utility within a single operating-region assessment. The contribution of this work is not the introduction of a new classifier architecture. Rather, it is the operational integration of probability reliability, threshold behavior, and net clinical benefit into the evaluation of 30-day readmission prediction. Specifically, the study applies this workflow to retrospective multi-site inpatient data from resource-constrained hospitals and examines whether model outputs remain interpretable and actionable when translated into discharge-planning thresholds. By doing so, the manuscript positions readmission prediction as a decision-support problem, not only as a classification task evaluated by ROC-AUC [17], [18], [25].

The remainder of this study is organized as follows. Section II reviews related work on readmission prediction, calibration, threshold policy, and decision-curve-based evaluation in clinical AI. Section III describes the study design, data source, model development pipeline, and evaluation protocol, including calibration assessment, threshold sensitivity analysis, and decision curve analysis. Section IV presents the results across discrimination, calibration, threshold behavior, and decision utility. Section V interprets the findings in relation to deployment readiness, threshold governance, model monitoring, and study limitations. Section VI summarizes the main conclusions and implications for accountable readmission prediction in resource-limited hospital settings.

II. Related Works

A. Readmission Prediction and EHR-Based Machine Learning

Thirty-day readmission prediction has been studied extensively because readmission is both a quality indicator and an operational concern for hospitals. Earlier work has shown that electronic health records can support risk prediction by combining demographic information, laboratory values, vital signs, prior utilization, and admission-related variables [1], [5], [6]. Within this literature, machine learning and deep learning models have often been evaluated mainly by discrimination metrics, especially ROC-AUC, because these metrics summarize how well patients are ranked by predicted risk [6], [9]. This ranking perspective is useful, but it does not fully answer the deployment question faced by discharge teams: whether a predicted probability can be trusted when deciding which patient should receive follow-up, counselling, or post-discharge monitoring.

Systematic reviews of readmission prediction models have also shown that methodological quality and validation practices vary considerably across studies [2], [3]. In many cases, models report acceptable or high discrimination, but provide limited evidence about probability reliability, threshold behavior, and clinical usefulness under realistic operating constraints. This creates a gap between retrospective model performance and workflow-level applicability. The present study is positioned within this gap. Rather than treating readmission prediction as a purely ranking problem, it evaluates whether the model outputs remain usable when translated into probability-based discharge decisions.

B. Calibration and Probability Reliability in Clinical AI

Calibration has become increasingly important in clinical prediction because risk scores are often interpreted as probabilities during decision-making [12], [13]. A model with good discrimination may still overestimate or underestimate risk in clinically important strata. In readmission prediction, this distinction matters because a patient assigned a risk of 0.20 may be handled differently from a patient assigned a risk of 0.08, even if both patients are ranked correctly relative to others. When calibration is poor, the same threshold can route too many low-risk patients into follow-up or fail to identify patients whose risk is clinically meaningful.

Recent clinical AI guidance has therefore emphasized that deployment-oriented validation should not stop at ROC-AUC

[17], [18], [25], [26], [27]. Probability reliability, lifecycle monitoring, and recalibration planning are part of responsible model use, particularly when predictions are used to allocate finite clinical resources. In practical terms, this means that calibration should be monitored through measures such as reliability diagrams, Brier score, expected calibration error, and calibration slope. The current study follows this direction by comparing the neural network and random forest not only by discrimination, but also by the degree to which their predicted probabilities align with observed 30-day readmission rates.

C. Threshold Policy, Decision Utility, and Operating Regions

A second limitation in much of the readmission prediction literature is the treatment of classification thresholds as technical defaults rather than clinical policy choices. The conventional 0.50 threshold is rarely aligned with discharge-planning priorities, because hospitals may prefer to identify more high-risk patients even if this increases the number of false positives [7], [19], [32]. The appropriate threshold depends on the relative harm of missed readmissions, the available follow-up capacity, and the workflow used to act on model outputs. For example, if a transitional care team can make 30 follow-up calls per week, then the threshold should be chosen with that capacity in mind rather than inherited from a generic binary classifier.

Decision Curve Analysis provides a way to evaluate whether model-guided action produces clinical benefit across threshold values [23], [24]. This is especially relevant in resource-limited hospitals, where the cost of false positives is not only statistical but operational: each flagged discharge may require staff time, documentation, and follow-up coordination. By combining threshold sensitivity analysis with decision curve analysis, the present study moves beyond the search for a single "best" threshold and instead identifies a clinically acceptable operating region. This region-based approach links model evaluation with implementation planning because it specifies where recall, false negative reduction, probability reliability, and net benefit remain jointly defensible.

D. Positioning of the Present Study

The present study contributes to this literature by integrating four evaluation domains into a single deployment-oriented workflow: discrimination, calibration, threshold sensitivity, and decision utility. Its contribution is not the proposal of a new classifier architecture, nor does it claim that one model family is universally superior across settings. Instead, the study demonstrates how a readmission model can be examined before implementation by asking four linked questions: whether it ranks patients accurately, whether its probabilities are reliable, how its threshold choices affect missed readmissions and workload, and whether model-guided action provides positive net benefit. This positioning is especially relevant for resource-constrained hospitals, where a threshold decision has concrete operational consequences, for example, the number of follow-up calls that a discharge team can complete during a working week.

III. Methods

A. Study Design and Data Source

This study used a retrospective, multi-centre design based on real-world inpatient records collected between 2022 and 2024 from two government hospitals operating under resource-limited conditions. The dataset was drawn from Rafidia Hospital and Al-Watani Hospital, subject to institutional approval and de-identification procedures. These hospitals represent a clinically relevant resource-constrained setting because the records include heterogeneous documentation practices, non-trivial missingness patterns, and operational constraints that commonly affect clinical data quality in low-resource environments [2], [8]. Accordingly, the dataset was suitable for evaluating internal deployment readiness under non-ideal data conditions, but it was not used to establish external generalizability across all resource-limited hospitals. Fig. 1 illustrates the multidimensional structure to assess the clinical deployability and operating-region identification.

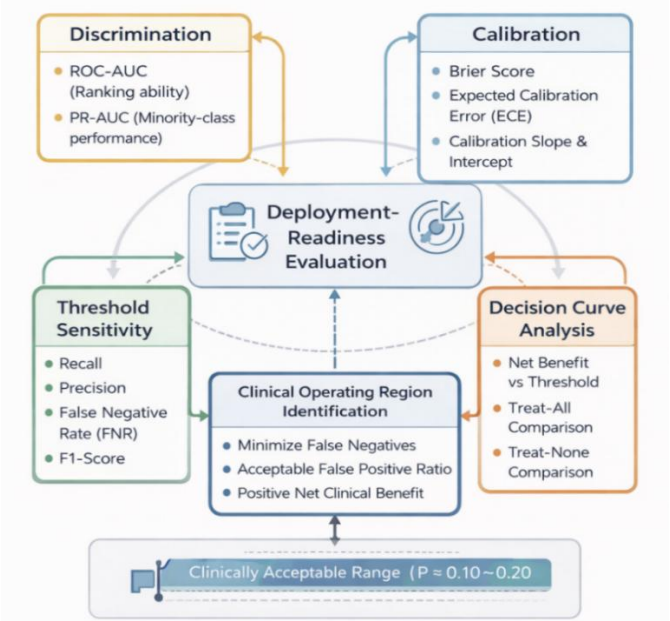

Fig. 1. Deployment-oriented clinical AI evaluation framework.

To minimise optimistic bias and prevent information leakage, we applied patient-level, group-aware data partitioning using unique patient identifiers (eid). A GroupShuffleSplit strategy was used to create non-overlapping training (70%), validation (15%), and held-out test (15%) subsets, with the constraint that no patient appeared in more than one subset. As a concrete operational safeguard, we programmatically verified that the intersection of patient identifiers across splits was empty and stored the split indices and random seed in an audit file so the exact partitions could be reproduced. All reported discrimination, calibration, Brier score, expected calibration error (ECE), and decision curve analysis (DCA) results were computed on the held-out test set only. Group-aware splitting is increasingly recommended in clinical machine learning to preserve validity and reduce inflated performance estimates that arise when repeated encounters from the same patient are inadvertently distributed across training and evaluation sets [8], [27]. This approach also aligns with contemporary guidance on validation practice for clinical prediction models [25].

No independent external validation cohort from a geographically or organizationally distinct institution was available for the present analysis. Therefore, all performance estimates should be interpreted as patient-level internal validation on a held-out test set rather than external validation. To reduce over-optimistic estimation within the available data, all preprocessing steps were fitted on the training set only, patient identifiers were kept non-overlapping across splits, and the final discrimination, calibration, threshold, and decision-curve results were calculated exclusively on the held-out test set.

The primary outcome was 30-day all-cause readmission following discharge. Analyses were conducted under institutional ethical approvals with de-identification procedures consistent with current reporting and governance expectations for clinical AI studies [17], [28]. See Table I.

TABLE I. COHORT CHARACTERISTICS AND DATA OVERVIEW

Category	Variable	Overall Cohort (N = 30,000)
Outcome distribution	30-day readmission, n (%)	4,500 (15.0%)
	No readmission, n (%)	25,500 (85.0%)
Demographics	Age, mean (SD), years	57.8 (18.4)
	Male sex, n (%)	15,960 (53.2%)
	Female sex, n (%)	14,040 (46.8%)
Admission characteristics	Length of stay, median (IQR), days	5 (3–8)
	Emergency admission, n (%)	18,750 (62.5%)
	Prior hospitalization (12 months), n (%)	9,840 (32.8%)
Feature composition	Demographic variables	6
	Laboratory variables	18
	Vital sign variables	6
	Admission-related variables	12
	Total features used in modeling	42
Missing data summary	Variables with <5% missingness	31 (73.8%)
	Variables with 5–20% missingness	9 (21.4%)
	Variables with >20% missingness	2 (4.8%)
	Imputation strategy	Median (continuous); Mode (categorical)
Data splitting strategy	Train set, n (%)	21,000 (70%)
	Validation set, n (%)	4,500 (15%)
	Test set, n (%)	4,500 (15%)
	Patient-level split	Applied
	Data leakage prevention	No patient overlap

B. Predictive Models

We evaluated two representative machine learning models: 1) a multilayer perceptron (MLP) neural network as the primary model, and 2) a random forest (RF) classifier as a baseline comparator. The purpose of this comparison was not to establish a full model leaderboard, but to examine how two commonly used nonlinear model families behave when evaluated through discrimination, calibration, threshold sensitivity, and decision utility under the same patient-level split and preprocessing protocol.

Neural network architecture and training. The MLP was implemented as a feed-forward network with three fully connected hidden layers (128, 64, and 32 units) using ReLU activations. Optimisation used Adam with an initial learning rate of 0.001 and L2 regularisation ($\alpha = 1 \times 10^{-4}$). Training proceeded for up to 300 iterations with early stopping based on validation performance, using a tolerance of 1×10^{-4} and patience of seven consecutive iterations without improvement. A fixed random seed (42) was used to support reproducibility.

A leakage-free, column-wise preprocessing pipeline was implemented for all data. Scaled, continuous variables were preprocessed with training set statistics only. Skewed continuous predictors (e.g., length of hospital stay and laboratory results) were standardized by median and inter-quartile ranges (IQR), whereas other continuous predictors were standardized by mean and standard deviation. Categorical predictors were treated equally during each split of the data, and missing data were replaced using medians for continuous variables and modes for categorical variables (see Table I). All preprocessing transformations were fit to the training set, and therefore, all subsequent transformations were performed without any additional modifications to the validation and testing sets, thereby eliminating potential feature-engineering-induced leakage from the training data. The Multilayer Perceptron (MLP) was chosen because it may be able to capture the non-linear relationships among the many, diverse demographic, laboratory, and admission-related variables common to EHR-based predictive tasks [1],[4],[6].

Random Forest Baseline Model. A Random Forest (RF) model was utilized as a baseline model because RF models are typically effective for structured tabular data, handle multiple types of predictor variables and have been extensively compared with other models within clinical predictive modeling applications [29] to allow comparison of the behavior of probabilities produced by both modelling families utilizing the same split of the data, same preprocessing pipeline, and same evaluation metrics. The objective of this research was not to achieve maximum discrimination via an exhaustive search of possible architectures and/or hyperparameters; rather, it was to evaluate how well the ML models were calibrated, sensitive to threshold decisions, and useful for making decisions within real-world operating conditions [18].

Benchmarking scope. Clinically established readmission scores such as LACE+ and HOSPITAL, as well as gradient-boosted tree models, are important comparators in broader readmission prediction research. However, the present study did not estimate these additional benchmarks. This decision reflects the study's primary aim, which was to demonstrate a deployment-oriented evaluation workflow rather than to optimize predictive performance across all possible model families. It also avoids introducing unreported model outputs or score reconstructions that were not part of the original analysis. Accordingly, the results should be interpreted as an internally validated comparison between NN and RF probability outputs, not as evidence that these two models exhaust the set of deployable alternatives for 30-day readmission prediction.

C. Multi-Dimensional Evaluation Framework

Consistent with current recommendations for evaluating prediction models intended for clinical use, assessment was structured across four complementary domains: discrimination, calibration, threshold sensitivity, and decision utility [3], [12], [17], [27]. This design reflects the position that discrimination alone is insufficient for clinical readiness because deployment decisions depend on probability validity and the consequences of acting on model outputs [25], [18].

Fig. 2 describes the model-training pipeline, probability generation, discrimination analysis, calibration assessment, threshold sensitivity evaluation, and decision curve analysis used to assess clinical reliability and operational suitability.

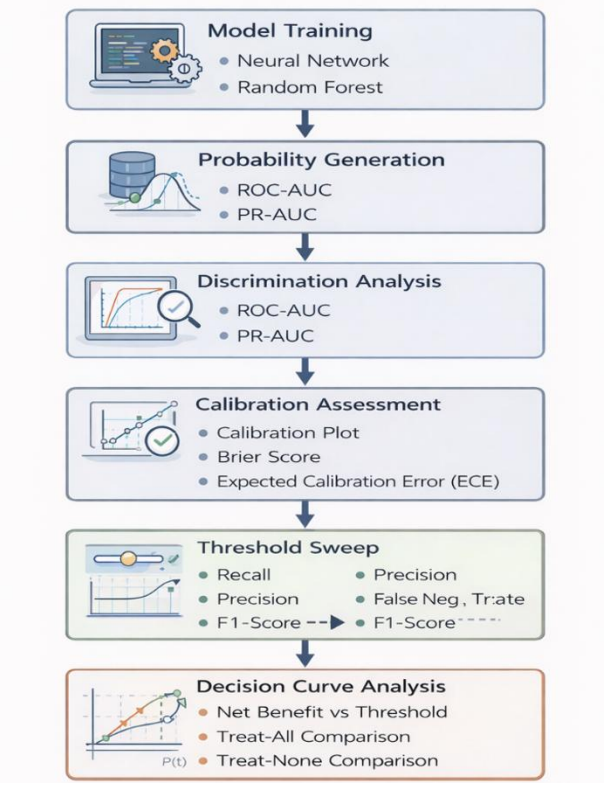

Fig. 2. Evaluation workflow for deployment-readiness assessment of clinical prediction models.

1) Discrimination assessment: Discrimination for both models was quantified using ROC-AUC, which summarises ranking performance across all possible thresholds. Because readmissions comprised approximately 15% of cases, we also evaluated precision–recall behaviour to characterise minority-class detection, and we report PR-AUC for the primary model as a complement to ROC-AUC under outcome imbalance [30], [31]. Discrimination was treated as necessary but not sufficient, and it was interpreted alongside calibration and decision-focused analyses.

2) Calibration assessment: Calibration analyses examined agreement between predicted probabilities and observed 30-day readmission rates. We computed reliability diagrams using observed versus predicted risk across deciles, the Brier score as an overall probabilistic accuracy measure, ECE as a bin-based summary of miscalibration, and calibration intercept and slope estimated via logistic recalibration [12], [27]. We also report the Hosmer–Lemeshow goodness-of-fit test as a conventional, though sample-size-sensitive, diagnostic. Calibration is repeatedly identified as a prerequisite for safe clinical deployment because it governs whether a predicted risk can be interpreted as an actionable probability rather than an uncalibrated score [12], [27]. Given outcome imbalance, we interpret the Brier score alongside complementary calibration measures, since prevalence can influence the absolute

magnitude of the Brier score [18]. ECE was computed using binned differences between predicted and observed event rates, consistent with common practice in probabilistic reliability evaluation [34]. For the recalibration sensitivity analysis, calibration metrics were reported separately for the original random forest, Platt-scaled random forest, and isotonic-regression-calibrated random forest, with all recalibration functions fitted on the validation set and evaluated only on the held-out test set. Reliability diagrams were generated by comparing observed and predicted 30-day readmission probabilities across risk deciles. These plots were used to visually assess departures from perfect calibration and are reported with the calibration results. Threshold Sensitivity Analysis.

Default probability thresholds, particularly 0.50, seldom reflect clinical risk tolerance or operational capacity in risk stratification tasks [7], [19], [32]. We therefore performed a systematic threshold sweep over probability cut-offs from 0.05 to 0.50 in increments of 0.01. At each threshold, we computed sensitivity (recall), precision, false negative rate (FNR), and F1-score (Table II). Because missed high-risk patients can undermine discharge safety pathways, the analysis emphasised FNR as a safety-relevant metric, while also documenting the corresponding workload implications captured by false positives [7], [34].

TABLE II. CONFUSION MATRIX PERFORMANCE METRICS AT DEFAULT (0.50) AND OPTIMIZED (0.12) PROBABILITY THRESHOLDS ON THE HELD-OUT TEST SET (N = 4,500)

Metric	Threshold = 0.50	Threshold = 0.12
True Positives (TP)	165	585
False Positives (FP)	151	1,018
True Negatives (TN)	3,674	2,807
False Negatives (FN)	510	90
Recall (Sensitivity)	0.244	0.867
Precision	0.522	0.365
F1-score	0.333	0.514
Overall Accuracy	0.853	0.754

3) Decision Curve Analysis (DCA): Clinical utility was evaluated using DCA across a range of probability thresholds [18], [19]. Net benefit was computed as:

$$\text{Net Benefit} = (TP / N) - (FP / N) \times (pt / (1 - pt)), \quad (1)$$

where, pt denotes the decision threshold. Model net benefit was compared against treat-all and treat-none strategies, and we also examined the default-threshold strategy to illustrate the impact of threshold choice on decision value. DCA provides an interpretable summary of whether decisions guided by model predictions improve outcomes under explicit trade-offs between false positives and false negatives, complementing discrimination and calibration analyses [24], [33]. Methodological guidance has increasingly positioned DCA as a core component of deployment-oriented validation, particularly when probability estimates are used to allocate finite interventions [18]. Fig. 3 illustrates net benefit across probability thresholds (0.05–0.50) compared with treat-all and treat-none strategies.

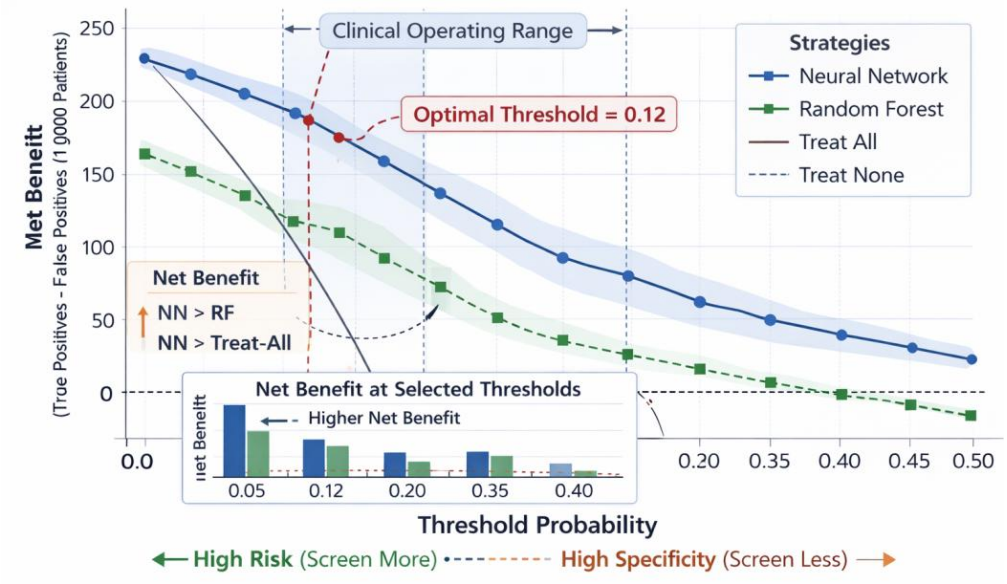

Fig. 3. Decision curve analysis of neural network and random forest models.

4) Clinical operating region identification: Rather than identify a "best" cut-off point for screening purposes based upon a single performance metric, we have identified a clinically acceptable operating region of thresholds which meet four criteria: improvement in the number of false negative results; an acceptable number of false positives to accommodate follow-up capacity, a positive net benefit in terms of DCA; and good calibration across all clinically relevant risk strata. The operating region approach is aligned with deployment-oriented expectations [16] [35] regarding consistency of decisions, safety, and fit into workflow under real-world constraints.

D. Statistical Analysis and Implementation

All analyses were implemented in Python using standard machine learning libraries within a reproducible computational environment. To quantify uncertainty, we used non-parametric bootstrap resampling on the held-out test set with 1,000 resamples drawn with replacement. For each resample, we recomputed ROC-AUC for both models and PR-AUC for the primary model, then recomputed calibration measures, including Brier score, ECE, calibration intercept, and calibration slope, and finally recomputed net benefit at selected thresholds for DCA. Confidence intervals were derived using the percentile method to obtain 95% intervals. This resampling-based approach provides a distributional view of metric variability, which is particularly relevant in imbalanced clinical prediction tasks and aligns with modern validation practice [23], [27]. Statistical significance was assessed at $\alpha = 0.05$ when hypothesis tests were reported, but interpretation prioritised effect magnitudes, interval estimates, probability reliability, and clinical interpretability over dichotomous p-value thresholds.

IV. Results

A. Discrimination Performance

Both models were highly discriminating at a global level on the test data for which they were partitioned using patient-level, group-aware partitioning. The Neural Network (NN) was able to achieve an ROC-AUC value of .868, showing that it ranked well in spite of high levels of imbalance. Using bootstrap resampling over 1000 iterations supports the stability of this estimate, and suggests that the ranking ability of the model was not due to a small number of samples in the test data.

A Random Forest (RF) Baseline performed better than the NN, with a higher ROC-AUC (.880), but also nearly identical precision – recall AUC (.460) for both models when looking at positive class detection, as readmission represented around 15% of the sample population. Overall, the discrimination results suggest that the models have similar risk rankings at a global level, although the RF has a slight advantage over the NN on ROC-AUC.

Discrimination, however, did not determine deployment readiness on its own. Because discharge workflows require binary decisions, and because these decisions depend on both threshold choice and probability reliability, subsequent analyses focused on threshold-dependent behaviour, calibration, and decision utility.

Fig. 4 shows the receiver operating characteristic curves used to summarize discrimination across thresholds. The precision–recall curve highlights positive-class detection under class imbalance, with readmissions accounting for approximately 15% of cases.

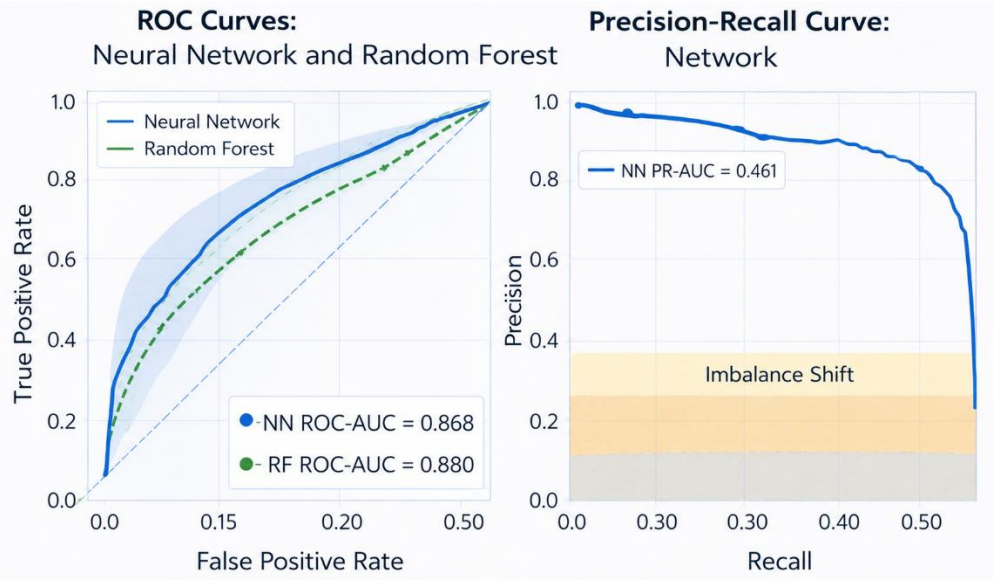

Fig. 4. ROC curves for the neural network and random forest models, and precision-recall curve for the neural network on the held-out test set.

Threshold-dependent behaviour showed why discrimination-only summaries were incomplete. At the default threshold of 0.50, recall for 30-day readmission was 0.244, meaning most readmissions were not flagged. When the decision threshold was set to 0.12 using a recall-oriented F-beta strategy ($\beta = 2$), recall increased to 0.867, with F1-score improving to 0.514 and overall accuracy equal to 0.754. This change reduced false negatives from 510 to 90 cases. It also changed operational workload: the number of flagged discharges increased from 316 at threshold 0.50 (TP + FP = 165

+ 151) to 1,603 at threshold 0.12 (TP + FP = 585 + 1,018), which corresponds to 7.0% versus 35.6% of the 4,500-patient test set (Table III). This illustrates that the same model can shift from under-detection to high-sensitivity screening depending on the operating threshold, with direct implications for follow-up capacity.

TABLE III. DISCRIMINATION AND CLASSIFICATION METRICS AT SELECTED THRESHOLDS (NEURAL NETWORK)

Metric	Threshold = 0.50 (Default)	Threshold = 0.12 (Optimized)
Recall (Sensitivity)	0.244	0.867
Precision	0.522	0.365
F1-score	0.333	0.514
Overall Accuracy	0.853	0.754
False Negatives (n)	510	90
False Negative Rate (FNR)	0.756	0.133

The comparison between thresholds 0.12 and 0.50, therefore, reflects a clinically meaningful trade-off. Lowering the threshold substantially reduced missed high-risk patients, while increasing false positives and the number of patients routed into a follow-up pathway. Although RF achieved marginally higher ROC-AUC, later sections show that probability reliability and decision-analytic performance differed between models, reinforcing that small differences in discrimination did not map cleanly onto deployment-relevant behaviour.

B. Calibration Assessment

Calibration analyses on the held-out test set examined whether predicted probabilities aligned with observed readmission frequencies. This matters because calibrated probabilities are interpretable as risk estimates, and interpretability is required for defensible threshold policies in discharge planning.

1) Brier score: A prevalence-only baseline that assigns every patient a 0.15 readmission risk achieved a Brier score of 0.1275. Both learned models improved substantially on this baseline. The NN produced a Brier score of 0.0950, and RF produced the same value, 0.0950, indicating comparable overall probabilistic accuracy when assessed by mean squared error of predicted probabilities. Bootstrap resampling supported the stability of these values. At the same time, the Brier score aggregates errors across the full probability range, so it can conceal differences in probability scaling and bin-level alignment, particularly in imbalanced datasets.

Probability density distributions of predicted risk scores for the neural network and random forest models evaluated exclusively on the patient-level held-out test set. Differences in spread and central tendency were examined further using calibration-specific metrics (see Fig. 5).

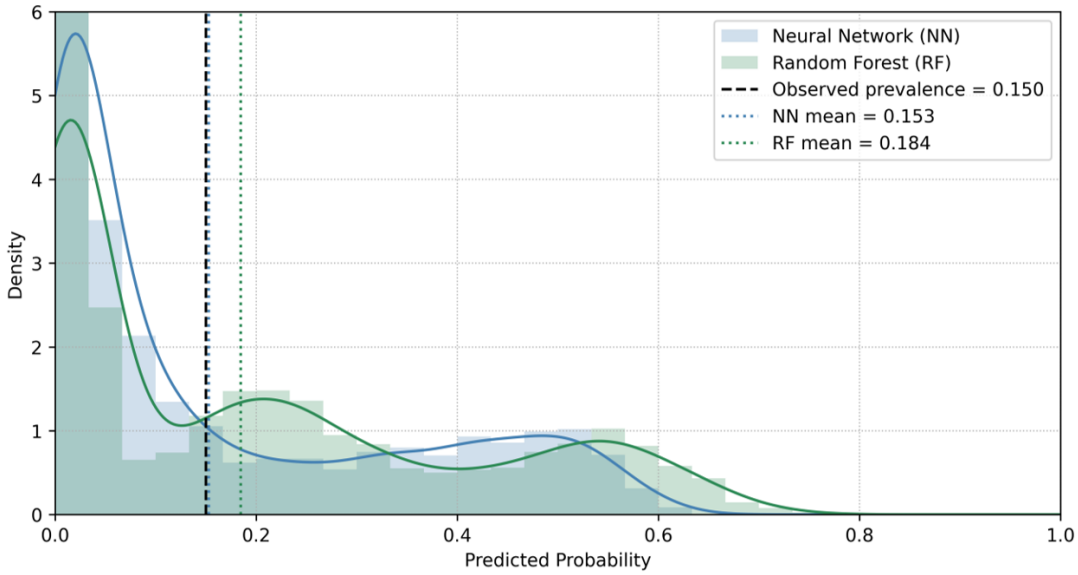

Fig. 5. Distribution of predicted 30-day readmission probabilities on the held-out test set.

2) Expected Calibration Error (ECE): Expected calibration error, computed on the held-out test set using equal-width bins, distinguished the models more clearly than the Brier score. The NN achieved ECE = 0.0128, whereas RF achieved ECE = 0.0348. The smaller value for NN indicates closer agreement between predicted probabilities and observed event rates across probability bins. In contrast, RF showed larger bin-level discrepancies, despite its marginal advantage in ROC-AUC. This contrast is important for clinical use because discharge decisions are often made in narrow probability bands, where bin-level miscalibration can shift which patients are flagged. Table IV shows the calibration analysis performed exclusively on the held-out test set. The results indicate that while both models have very similar total accuracy, the Neural Networks are capable of producing much more reliable probability estimates than Random Forests. Probability estimates are critical for clinical decision support, and therefore, this finding is highly relevant to our investigation.

TABLE IV. CALIBRATION PERFORMANCE ON THE HELD-OUT TEST SET

Metric	Neural Network (NN)	Random Forest (RF)	Ideal Value
Brier Score	0.095	0.095	0
Expected Calibration Error (ECE)	0.013	0.035	0
Calibration Intercept	−0.028	−0.216	0
Calibration Slope	0.971 (95% CI: 0.94–1.00)	1.202	1

3) Calibration intercept and slope: Recalibrating the logistic model to the held-out test data produced additional evidence as to how systematically biased and/or distorted the probability scales were for each model. The NN had an intercept of -0.028 and a slope of 0.971 (95% CI: 0.94-1.00). This demonstrates little systematic bias and appropriate probability scale distortion across risk levels. In contrast, the RF model's intercept was -0.216, and its slope was 1.202. This suggests that the RF model has larger deviations from ideal calibration and a larger probability scale distortion. Overall, these findings show that similar discrimination and Brier scores can conceal meaningful differences in uncalibrated probability reliability. The random forest showed larger deviation in ECE and calibration slope than the neural network, but this result should not be interpreted as evidence that the random forest model family is inherently less suitable for deployment. Rather, it indicates that the RF probabilities, as evaluated in their uncalibrated form, would require post-hoc recalibration assessment before any deployment-oriented comparison could be considered final.

This is evident in Fig. 6, where we show the observed versus the predicted 30-day readmission probability estimates for each risk level. The diagonal reference line represents perfect calibration, and the closer a model falls to this line, the more reliable its estimated probabilities are.

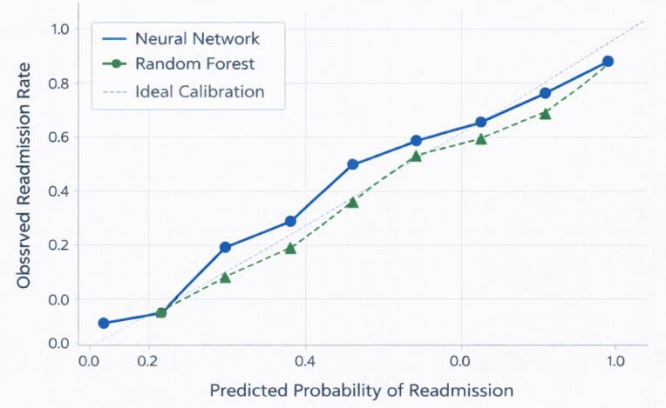

Fig. 6. Calibration curves (reliability diagrams) for neural network and random forest models.

C. Threshold Sensitivity Analysis

The results of the threshold sensitivity analyses are shown in Fig. 7. They indicate the size of the clinical performance variations associated with varying the decision threshold. For example, at the default threshold of 0.50, the model had a recall of 0.244. At the default threshold of 0.50, the model operated as a high-specificity but low-sensitivity rule. It identified only 24.4% of readmitted patients and therefore missed most patients who later experienced 30-day readmission. In contrast, when the threshold was set to 0.12, the recall increased to 0.867, the number of false negatives decreased from 510 to 90, which is roughly a sixfold decrease in missed readmissions. These data demonstrate that the conventional default thresholds used for decision-making in this context were inappropriate for achieving safe levels of care and further illustrate why the selection of the threshold should be viewed as a clinical decision based on available resources/workflow capacity rather than as an implicit statistical default.

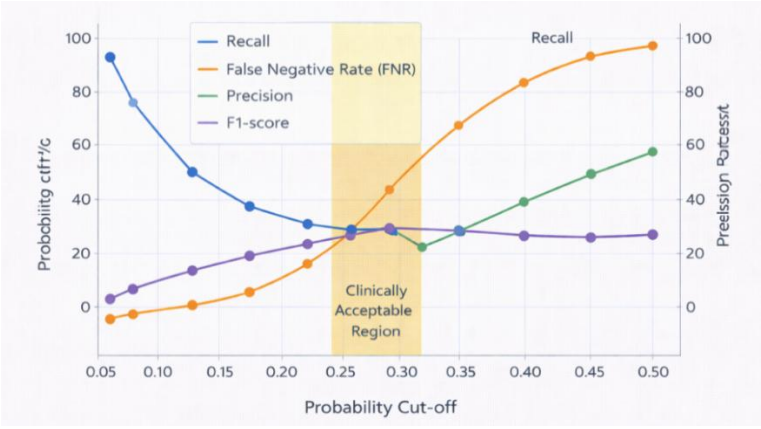

Fig. 7. Threshold sweep analysis for the neural network model.

Fig. 7 compares the recall, false negative rate (FNR), precision, and F1 score across different probability cut-off values ranging from 0.05 to 0.50. The shaded area (0.10–0.20) represents the clinically acceptable operating range, where there is high recall, a significant number of false negatives are eliminated, and sufficient decision utility is maintained.

D. Decision Curve Analysis

Decision Curve Analysis (DCA) examined whether taking action based on model predictions would result in net benefit over a wide range of thresholds. Across clinically relevant thresholds of 0.05 to 0.25, NN was found to have a consistent favorable net benefit when compared to treat-all and treat-none strategies. Using bootstrap resampling at an optimized threshold of 0.12, DCA produced a mean net benefit of 0.097 with a 95 percent confidence interval (CI): 0.093–0.101 that reflected relatively stable decision utility in the sensitivity-oriented operating range used for discharge screening.

Fig. 8 compares the net benefit across threshold probabilities for the NN strategy with the treat-all and treat-none strategies. NN has a favorable and relatively stable net benefit in the clinically relevant threshold ranges.

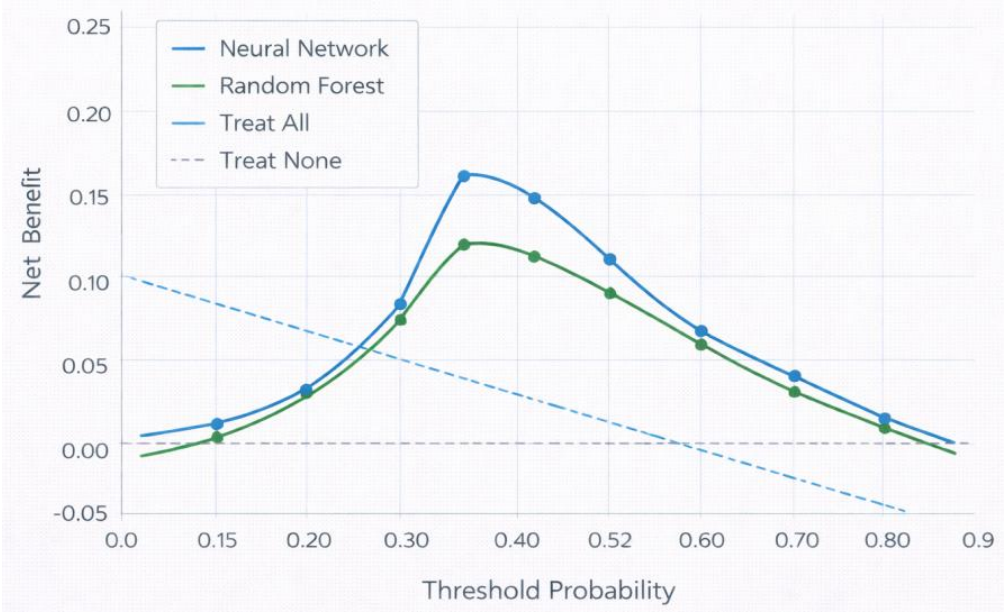

Fig. 8. Decision curve analysis comparing NN and RF strategies.

Although RF did show some positive net benefits at certain thresholds, it should be interpreted with caution, given the calibration distortions discussed previously. As discussed previously, these distortions may produce apparent shifts in gains across thresholds as a function of probability scaling distortion. Conversely, the combination of strong calibration and positive net benefits supports the suitability of NN for use in discharge planning workflows that rely on probability estimates to inform follow-up priority.

E. Clinical Operating Region Identification

The most appropriate deployment region, where the model would both correctly identify the individuals who needed treatment and provide enough information about how many people were treated, was determined by combining discrimination, calibration, threshold sensitivity, and decision utility instead of relying solely on one or more of those statistics to find the "best" threshold. The acceptable deployment region for clinical use was defined by four performance criteria: Recall should be at least 0.80; Calibration Error should be less than 0.02; Calibration Slope should be between 0.95 and 1.05 to allow the preservation of probability scaling; and there should be a positive Net Benefit relative to treating all and treating none. Using these criteria, thresholds ranging from 0.10 to 0.20 were found to be the best deployment region for NN. Within this range, NN provided good Recall (0.80 – 0.87), small Calibration Error (ECE = 0.013), and nearly perfect Probability Scaling (Slope = 0.971), with a positive Net Benefit. Therefore, the deployment region identified is a practical trade-off between significantly reducing False Negatives while maintaining Positive Decision Utility, see Table V.

TABLE V. INTEGRATED PERFORMANCE SUMMARY WITHIN THE CLINICALLY ACCEPTABLE OPERATING REGION (0.10–0.20)

Evaluation Dimension	Metric	Observed Performance (NN)	Interpretation Within Operating Region
Discrimination	ROC-AUC	0.868	Adequate global ranking performance
	PR-AUC	0.461	Stable minority-class detection under class imbalance
Calibration	Expected Calibration Error (ECE)	0.013	Minimal deviation between predicted and observed probabilities

	Calibration Slope	0.971 (95% CI: 0.94–1.00)	Probability scaling is closely aligned with the ideal (≈1)
Threshold Sensitivity	Recall	0.80–0.87	Substantial improvement in high-risk patient detection
	False Negatives	Reduced from 510 to 90 cases	Marked reduction in under-detection relative to default threshold (0.50)
Decision Utility	Net Benefit (DCA)	Positive vs. treat-all and treat-none strategies	Favorable clinical decision utility within operating range

V. Discussion

A. Beyond Discrimination: Why ROC-AUC Does Not Establish Deployment Readiness

This work supports the idea that it is possible for high discrimination and a number of potential deployment issues to be present together. The neural network produced a ROC-AUC of 0.868, while the random forest model generated a slightly better result at 0.880; however, neither of the global summary measures of rankings provided evidence that the probability predictions made by either model would be accurate enough to inform threshold-based discharge decisions. Discrimination, which is the focus of both of these models, addresses how well the model can rank patients in order of their predicted risk level, but does not provide information on the quantitative accuracy of the risk estimates made by the model to be used in actual decision-making processes (e.g., who should be discharged next) [12], [3], [27]. Therefore, there are numerous examples in the literature that have warned against using discrimination as the sole measure of clinical suitability for AI systems [12], [3], [27] and that other factors such as calibration stability, threshold alignment, and post-deployment monitoring and maintenance will need to be addressed to ensure trustworthy and lifecycle aware clinical AI systems [17], [16]. The findings reported here are consistent with this direction, suggesting that simply knowing the degree of discrimination of an AI system is not sufficient to determine if the system may be safely deployed and operated within real-world operational constraints.

B. Calibration Reliability Under Outcome Imbalance

The differences between the two predictive models were more apparent in calibration than in discrimination. Although both models produced the same Brier score (0.095) and similar discrimination, the calibration metrics indicated different levels of probability reliability under the present evaluation protocol. The uncalibrated random forest showed a higher expected calibration error (ECE = 0.035) and a calibration slope of 1.202, suggesting probability-scale distortion across risk strata. The neural network showed lower ECE (0.013) and a calibration slope closer to the ideal value (0.971; 95% CI: 0.94–1.00), indicating closer alignment between predicted and observed risk in this held-out test set. However, because post-hoc recalibration was not performed, these results should be interpreted as a comparison of the evaluated probability outputs, not as evidence that the random forest model family is inherently less suitable for deployment.

This distinction is particularly important when making clinical predictions using an imbalance of events. Global summaries such as the Brier Score pool squared error across all probability ranges, and since the event rate for this study was roughly 15%, the majority of errors in the decision-making space may have been diluted due to the large number of members in the non-event class. In contrast, ECE and the recalibration slope provide visibility to the reliability of the predictive probabilities in those bins/strata that are operationally relevant to the workflow, i.e., the decision thresholds [12], [18], [36].

Therefore, contemporary practice recommends assessing calibration as part of evaluating how well a model has been deployed, rather than treating it as simply another metric to report on discrimination performance [12], [3], [18]. The probability reliability of predictive modeling is not merely a cosmetic quality for hospital readmission workflows, but impacts who will be flagged, what clinicians interpret as the predictive risk means, and if the same threshold remains consistent in behavior over time.

C. Threshold Optimisation as a Clinical Design Decision

Beginning with Threshold Analysis, we see why threshold selection is considered part of both Clinical Design/Governance and not Conventional Practice. With a default threshold value of .50, we can see that Recall is .244, meaning there would have been a significant amount of readmission of patients that would have been missed using the Default Rule. Using an F-Beta Strategy (beta = 2) for a Recall-Oriented Approach, we can adjust the threshold to .12, and our Recall goes up to .867 while False Negatives drop to 90, down from 510.

The clinical implications are direct; a hospital discharge process that relies on identifying and addressing potential issues early in a patient's discharge will not perform as intended when the early identification tool misses three out of four readmissions.

The trade-off is also direct; to achieve a Recall of .867 using the Recall-Oriented Approach at the lower threshold (.12), we had to increase the number of discharges that were "flagged" to 1603 out of 4500 in the Test Set (True Positives + False Positives) compared to 316 at the higher threshold value of .50. This means that for every patient that has to undergo additional follow-up calls, and/or additional staffing capacity due to the need to schedule additional time for follow-up calls, the hospital will incur the associated cost. For this reason, recent literature recommends making the threshold selection based upon the specific trade-offs relevant to the deployment, documenting the reasoning behind the selection of the threshold, and matching the operating point of the system to the service capacity and level of risk that the organization can tolerate, rather than simply applying a generic default cut-off [25],[7],[37]. Our study provides empirical support for these recommendations: threshold selection affects the performance of systems that produce clinically important decisions; therefore, threshold

selection should be viewed as part of the design of a deployment, and thus should be subject to clearly defined escalation processes and audit trails for change control, rather than being treated as post-hoc tuning [38].

D. Decision Curve Analysis and Actionable Utility

Decision Curve Analysis (DCA) enhanced calibration and threshold sweep by asking a deployment-based question, i.e., does the model led to positive net benefits when using the model as a basis for action, across plausible thresholds [23] [24]. The Neural Network maintained favorable net benefits compared to the "treat all" and "treat none" strategies across the entire range of thresholds used for sensitivity-based screening of patients (i.e., approximately 0.05-0.25). Net Benefit was most stable at the optimal threshold of 0.12; Mean=0.097 (95% CI: 0.093-0.101) under bootstrap resampling. The stability of net benefits is important because it suggests that the observed decision advantage was not due to an isolated sample-specificity.

The importance of net benefit in this context stems from its ability to explicitly represent the trade-offs between false positives and false negatives through the threshold odds term pt/(1-pt) [24] [33]; and link those trade-offs to clinical actions [24] [33]. There is an increasing recognition of DCA as being complementary to calibration [25] [18] since a model can be well calibrated but still have limited utility in making decisions based on the model's output if the decision thresholds are not aligned with clinical practices; and similarly, a model may demonstrate high discrimination yet perform poorly in decision-making terms with respect to the ranges of thresholds that are clinically meaningful [23].

The inclusion of DCA with both calibration and threshold sensitivity allowed for a shift in how models were interpreted from global measures of performance to actionable outcomes, which are the appropriate standards of evaluation for settings where models will be deployed.

E. Implications for Resource-Constrained Healthcare Settings

The impact of prediction error on operations within resource-limited settings is amplified. Due to staffing limitations, variability in the practice of documenting, and limited follow-up options, false positives will continue to over saturate services while False Negatives will result in avoidable deterioration before receiving appropriate support [39]. As such, probability reliability becomes a necessity for workflows in resource-constrained settings. In addition, if predictions regarding patient risk are consistently skewed, clinicians may have difficulty developing consistent priorities based upon the recommendations of the clinical decision support system, and may lose confidence in the clinical decision support system's recommendations as they appear to contradict their own clinical impressions. Therefore, lifecycle-oriented clinical AI models address risk communication (calibrated), monitoring (ongoing), and governance processes that define drift and the need to recalibrate as normal occurrences, rather than as unusual events [17][16].

In addition, under the present uncalibrated evaluation protocol, the neural network showed favorable probability reliability and positive decision utility in the 0.10–0.20 operating region. This supports its use as the primary model for interpreting the current internal validation results. It should not, however, be read as a categorical deployment preference over the random forest, because RF recalibration using Platt scaling or isotonic regression was not tested. A practical implication of this approach is that threshold adjustments must be paired with lightweight monitoring. For example, a monthly report could document the current version of the model, the active threshold(s), the total number of discharges that were flagged by the model, the observed readmission rate among the flagged patients, and a decile-based calibration summary. This would allow for the tracking of threshold adjustments and performance drift through an audit trail. The incorporation of routine monitoring provides an ability to connect the behavior of the model to the operational decisions made as a result of those behaviors and allows for the safe iterative refinement of the model.

F. A Structured Framework for Deployment Readiness

Beyond evaluating models together, this study also provides a structure for evaluation to make deployability measurable. The central idea is to treat readiness as evidence across linked domains rather than as a single score. Discrimination produces a summary of ranking performance; calibration establishes whether the model's probability estimates are valid; threshold sensitivity describes how safety and workload will shift as policies change; and DCA quantifies whether an action improves outcomes in the range of thresholds that a workflow can sustain. The additional step of identifying a clinically acceptable operating region translates these metrics into a bounded set of thresholds where performance and utility are jointly defendable. Framed in this way, evaluation supports not only model selection but also planning for operation, including threshold governance, escalation rules, and monitoring triggers, which are required to deploy in safety-sensitive domains [40], [41]. While readmission prediction is focused upon here, the same logic applies to other imbalanced clinical tasks where small shifts in threshold policy can materially change patient routing.

G. Methodological Contribution

The research findings provide a deployment-focused interpretation of model performance, which may be summarized as three interrelated claims. Firstly, discrimination alone was not sufficient to establish deployability; the models were judged to have similar rankings for risk but different levels of probability reliability and decision utility. Secondly, having equal Brier scores did not mean the two models had the same level of calibration; both ECE and slope showed clinically important differences in how they scaled their probabilities. Finally, choosing the threshold to use changed the safety and workload implications of deploying the model, leading to threshold governance and decision-utility assessment being central to validating the model rather than optional additions to it. These findings collectively reframed model evaluation from "Is the model correct?" to "Do the probabilities generated by the model provide reliable information, does the decision policy used with the model behave consistently with expectations over the intended operating domain, and do decisions based upon the output of the model result in improved outcomes when the

model is deployed in a real-world environment subject to realistic constraints?".

H. Limitations and Future Directions

Several limitations should be considered when interpreting the findings of this study. First, although the dataset included 30,000 inpatient records from two resource-limited government hospitals, the evaluation remains retrospective and internally validated. No independent external cohort from a geographically or organizationally distinct hospital system was available. Therefore, the findings support the usefulness of the proposed evaluation workflow within the tested setting, but they should not be interpreted as evidence that the same model, threshold range, or calibration behavior will transfer unchanged to other resource-limited hospitals. External validation is required before broader implementation, ideally using a separate hospital network with different documentation practices, admission profiles, case mix, and follow-up capacity.

Second, the current evaluation assessed decision utility retrospectively rather than prospectively measuring the effect of model-guided intervention in routine discharge workflows. A prospective study would provide stronger evidence by examining operational outcomes after intervention allocation, such as follow-up completion rate, clinician adherence to risk alerts, readmission reduction, escalation decisions, and patient-level safety outcomes. For example, if a threshold of 0.12 routes approximately one third of discharged patients into a follow-up pathway, the clinical value of this threshold should be tested against the actual number of calls, appointments, or case-management reviews that staff can complete during a defined working week.

Third, the random forest model was evaluated without post-hoc recalibration using methods such as Platt scaling or isotonic regression. This limits the interpretation of the calibration comparison because part of the observed calibration gap may be correctable through recalibration. Future work should therefore evaluate whether recalibrated random forest probabilities change the model comparison in terms of ECE, calibration slope, threshold behavior, and net benefit.

Fourth, the comparative evaluation was limited to a neural network and a random forest. The study did not benchmark clinically established readmission scores such as LACE+ or HOSPITAL, nor did it include gradient-boosted tree models. This limits claims about comparative predictive superiority. Future studies should evaluate the same deployment-oriented workflow across a broader benchmark set, including clinical scores, random forest, neural networks, and gradient-boosted methods, using identical patient-level splits, preprocessing rules, calibration metrics, threshold sweeps, and decision curve analysis. This would allow the operating-region framework to be assessed as a general evaluation procedure rather than as a comparison restricted to two model families.

Finally, future studies should explicitly examine distribution shift and calibration drift after deployment. This requires continuous monitoring, predefined triggers for recalibration, and governance procedures for documenting changes in thresholds and model versions over time [16]. Such procedures would support the transition from one-time model validation to lifecycle management, which is necessary for responsible clinical AI implementation.

VI. Conclusions and Future Directions

This study shows that strong discrimination is necessary but not sufficient for evaluating clinical prediction models intended for discharge-planning decisions. ROC-AUC can indicate whether a model ranks patients correctly, but deployment also requires calibrated probabilities, transparent threshold policies, and evidence of decision utility within a clinically workable operating range.

The results illustrate this point clearly. Although the neural network and random forest showed broadly similar global performance, their calibration behavior differed. The random forest achieved slightly higher discrimination, but it also showed larger calibration error and greater probability-scale distortion. By contrast, the neural network produced more reliable probability estimates, behaved more predictably across threshold changes, and maintained positive net benefit within the clinically reasonable operating region. These findings indicate that practical model selection should consider probability reliability and threshold behavior, not only small gains in discrimination.

By combining discrimination, calibration assessment, threshold sensitivity analysis, and decision curve analysis, this study provides a deployment-oriented evaluation workflow for 30-day readmission prediction. In practical use, this means that the selected threshold should be linked to follow-up capacity, documented as a standing decision rule, and monitored through an audit trail that records the model version, active threshold, flagged-discharge volume, and monthly calibration summary.

These conclusions should be interpreted within the study design. The analysis used retrospective inpatient data from two resource-limited government hospitals and relied on patient-level internal validation with non-overlapping training, validation, and held-out test partitions. Therefore, the results support the value of the proposed workflow in the tested setting, but they do not establish that the same model, threshold range, calibration behavior, or model preference will transfer unchanged to other hospitals or to other benchmark models.

Future work should include independent external validation and prospective workflow evaluation. Important implementation outcomes include follow-up completion rate, clinician response to risk alerts, escalation decisions, readmission outcomes, staff workload, and calibration drift over time. Future studies should also evaluate recalibration strategies, monitor distribution shift, and define governance rules for updating thresholds and model versions during clinical use.

Authors' Contributions

The authors' contributions are as follows: "Conceptualization, Samer Asad Malalha; methodology, Samer Asad Malalha; software, Ma Burhanuddin; validation, Hatem T M Duhair; formal analysis, Jamil Abedalrahim Jamil Alsayaydeh; investigation, Samer Asad Malalha; resources, Mazen Farid; data curation, Hatem T M Duhair; writing—original draft preparation,, Jamil Abedalrahim Jamil Alsayaydeh, Hatem T M Duhair and Mazen Farid; writing—

review and editing, Ma Burhanuddin and Mazen Farid; visualization, Samer Asad Malalha; supervision, Ma Burhanuddin; project administration, Samer Asad Malalha; funding acquisition, Mazen Farid. All authors have read and agreed to the published version of the manuscript." The relevant terms are explained in the CRediT taxonomy.

ACKNOWLEDGMENT

The authors express their gratitude to the Centre for Research and Innovation Management (CRIM) at Universiti Teknikal Malaysia Melaka (UTeM) for their valuable support in this research.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

The study was approved by the Clinical Research Center of An-Najah National University Hospital (Approval No. CRC_2025_0554; August 19, 2025). Approval to access anonymized patient records from government hospitals was also granted by the Palestinian Ministry of Health (Ref. No. 162/2161/2025; September 1, 2025), covering Rafidia Hospital and Al-Watani Hospital. All procedures were conducted in accordance with institutional and national ethical guidelines. Patient consent was waived due to the retrospective nature of the study and the use of fully anonymized electronic health records.

DATA AVAILABILITY

Data will be available on request.

REFERENCES

[1] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, Apr. 2019, doi: 10.1056/NEJMra1814259.

[2] B. Van Grootven, P. Jepma, C. Rijpkema, L. Verweij, M. Leeflang, J. Daams, M. Deschodt, K. Milisen, J. Flamaing, and B. Buurman, "Prediction models for hospital readmissions in patients with heart disease: A systematic review and meta-analysis," BMJ Open, vol. 11, no. 8, e047576, Aug. 2021, doi: 10.1136/bmjopen-2020-047576.

[3] L. Wynants et al., "Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal," BMJ, vol. 369, m1328, Apr. 2020, doi: 10.1136/bmj.m1328.

[4] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, A. K. M. Z. Hossain, V. Shkarupylo, and J. Pusppanathan, "The experimental studies of the automatic control methods of magnetic separators performance by magnetic product," ARPN Journal of Engineering and Applied Sciences, vol. 15, no. 7, pp. 922-927, 2020.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[6] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, Sep. 2018, doi: 10.1109/JBHI.2017.2767063.

[7] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," BMC Medicine, vol. 17, no. 1, p. 195, Oct. 2019, doi: 10.1186/s12916-019-1426-2.

[8] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nature Machine Intelligence, vol. 3, no. 3, pp. 199–217, Mar. 2021, doi: 10.1038/s42256-021-00307-0.

[9] S.-M. Zhou, R. A. Lyons, M. A. Rahman, A. Holborow, and S. Brophy, "Predicting hospital readmission for campylobacteriosis from electronic health records: A machine learning and text mining perspective," Journal of Personalized Medicine, vol. 12, no. 1, p. 86, Jan. 2022, doi: 10.3390/jpm12010086.

[10] J. A. J. Alsayaydeh, W. A. Indra, A. W. Y. Khang, V. Shkarupylo, and D. A. P. P. Jkatisan, "Development of vehicle ignition using fingerprint," ARPN Journal of Engineering and Applied Sciences, vol. 14, no. 23, pp. 4045–4053, 2019. [Online]. Available: http://www.arpnjournals.org/jeas/research_papers/rp_2019/jeas_1219_8024.pdf.

[11] B. Van Calster, L. Wynants, D. Timmerman, E. W. Steyerberg, and G. S. Collins, "Predictive analytics in health care: How can we know it works?" Journal of the American Medical Informatics Association, vol. 26, no. 12, pp. 1651–1654, Dec. 2019, doi: 10.1093/jamia/ocz130.

[12] M. N. Al-Andoli, Irianto, J. A. Alsayaydeh, I. M. Alwayle, C. K. N. Che Ku Mohd and F. Abuhoureyah, "Robust Overlapping Community Detection in Complex Networks With Graph Convolutional Networks and Fuzzy C-Means," in IEEE Access, vol. 12, pp. 70129-70145, 2024, doi: 10.1109/ACCESS.2024.3399883.

[13] R. D. Riley, K. I. E. Snell, J. Ensor, D. L. Burke, F. E. Harrell Jr., K. G. Moons, and G. S. Collins, "Minimum sample size for developing a multivariable prediction model: PART II, binary and time-to-event outcomes," Statistics in Medicine, vol. 38, no. 7, pp. 1276–1296, Mar. 2019, doi: 10.1002/sim.7992.

[14] C. G. Walsh, K. Sharman, and G. Hripcsak, "Beyond discrimination: A comparison of calibration methods and clinical usefulness of predictive models of readmission risk," Journal of Biomedical Informatics, vol. 76, pp. 9–18, Dec. 2017, doi: 10.1016/j.jbi.2017.10.008.

[15] J. A. J. Alsayaydeh, W. A. Y. Khang, W. A. Indra, V. Shkarupylo and J. Jayasundar. 2019. Development of smart dustbin by using apps. ARPN Journal of Engineering and Applied Sciences. 14(21): 37033711.

[16] K. Lekadir et al., "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," BMJ, vol. 388, e081554, Feb. 2025, doi: 10.1136/bmj-2024-081554.

[17] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement," BMJ, vol. 350, g7594, Jan. 2015, doi: 10.1136/bmj.g7594.

[18] B. Van Calster et al., "Evaluation of performance measures in predictive artificial intelligence models to support medical decisions: Overview and guidance," The Lancet Digital Health, vol. 7, no. 12, p. 100916, Dec. 2025, doi: 10.1016/j.landig.2025.100916.

[19] D. J. Hand, "Measuring classifier performance: A coherent alternative to the area under the ROC curve," Machine Learning, vol. 77, no. 1, pp. 103–123, Oct. 2009, doi: 10.1007/s10994-009-5119-5.

[20] K. B. Johnson et al., "Precision medicine, AI, and the future of personalized health care," Clinical and Translational Science, vol. 14, no. 1, pp. 86–93, Jan. 2021, doi: 10.1111/cts.12884.

[21] D. Piovani, R. Sokou, A. G. Tsantes, A. S. Vitello, and S. Bonovas, "Optimizing clinical decision making with decision curve analysis: Insights for clinical investigators," Healthcare, vol. 11, no. 16, p. 2244, Aug. 2023, doi: 10.3390/healthcare11162244.

[22] J. A. Alsayaydeh, M. Nj, S. N. Syed, A. W. Yoon, W. A. Indra, V. Shkarupylo and C. Pellipus, "Homes appliances control using bluetooth," ARPN Journal of Engineering and Applied Sciences, vol. 14 (19), pp. 3344-3357, 2019.

[23] J. Benitez-Aurioles, L. Wynants, N. Peek, P. Goodley, P. Crosbie, and M. Sperrin, "The continuous net benefit: Assessing the clinical utility of prediction models when informing a continuum of decisions," Diagnostic and Prognostic Research, vol. 10, no. 1, p. 8, 2026, doi: 10.1186/s41512-026-00224-z.

[24] A. J. Vickers, B. Van Calster, and E. W. Steyerberg, "Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests," BMJ, vol. 352, i6, Jan. 2016, doi: 10.1136/bmj.i6.

[25] O. Efthimiou, M. Seo, K. Chalkou, T. Debray, M. Egger, and G. Salanti, "Developing clinical prediction models: A step-by-step guide," BMJ, vol. 386, e078276, 2024, doi: 10.1136/bmj-2023-078276.

[26] A. Subbaswamy and S. Saria, "From development to deployment: Dataset shift, causality, and shift-stable models in health AI," Biostatistics, p. kxz041, 2019, doi: 10.1093/biostatistics/kxz041.

[27] R. D. Riley et al., “Calculating the sample size required for developing a clinical prediction model,” BMJ, p. m441, 2020, doi: 10.1136/bmj.m441.

[28] World Health Organization, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance, 1st ed. Geneva, Switzerland: World Health Organization, 2021.

[29] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[30] E. Richardson, R. Trevizani, J. A. Greenbaum, H. Carter, M. Nielsen, and B. Peters, “The receiver operating characteristic curve accurately assesses imbalanced datasets,” Patterns, vol. 5, no. 6, p. 100994, Jun. 2024, doi: 10.1016/j.patter.2024.100994.

[31] J. A. J. Alsayaydeh, Irianto, A. Aziz, C. K. Xin, A. K. M. Z. Hossain and S. G. Herawan, “Face Recognition System Design and Implementation using Neural Networks” International Journal of Advanced Computer Science and Applications(IJACSA), 13(6), June 2022, pp. 519-526. http://dx.doi.org/10.14569/IJACSA.2022.0130663.

[32] M. Pepe, The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford, United Kingdom: Oxford University Press, 2003.

[33] M. J. Pencina, R. B. D’Agostino, and E. W. Steyerberg, “Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers,” Statistics in Medicine, vol. 30, no. 1, pp. 11–21, Jan. 2011, doi: 10.1002/sim.4085.

[34] F. Futami and M. Fujisawa, “Information-theoretic generalization analysis for expected calibration error,” arXiv:2405.15709, 2025, doi: 10.48550/arXiv.2405.15709.

[35] M. Imani, M. Joudaki, A. Bagheri, and H. R. Arabnia, “Why ROC-AUC is misleading for highly imbalanced data: In-depth evaluation of MCC, F2-score, H-measure, and AUC-based metrics across diverse classifiers,” Technologies, vol. 14, no. 1, p. 54, 2026, doi: 10.3390/technologies14010054.

[36] E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd ed. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-16399-0.

[37] M. P. Sendak et al., “Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study,” JMIR Medical Informatics, vol. 8, no. 7, e15182, Jul. 2020, doi: 10.2196/15182.

[38] J. Feng et al., “Clinical artificial intelligence quality improvement: Toward continual monitoring and updating of AI algorithms in healthcare,” npj Digital Medicine, vol. 5, no. 1, p. 66, May 2022, doi: 10.1038/s41746-022-00611-y.

[39] S. E. Davis, R. A. Greevy, T. A. Lasko, C. G. Walsh, and M. E. Matheny, “Detection of calibration drift in clinical prediction models to inform model updating,” Journal of Biomedical Informatics, vol. 112, p. 103611, Dec. 2020, doi: 10.1016/j.jbi.2020.103611.

[40] J. A. J. Alsayaydeh, Irianto, M. F. Ali, M. N. M. Al-Andoli and S. G. Herawan, "Improving the Robustness of IoT-Powered Smart City Applications Through Service-Reliant Application Authentication Technique," in IEEE Access, vol. 12, pp. 19405-19417, 2024, doi: 10.1109/ACCESS.2024.3361407.

[41] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, “Adversarial attacks on medical machine learning,” Science, vol. 363, no. 6433, pp. 1287–1289, Mar. 2019, doi: 10.1126/science.aaw4399.