Journal of Science, Innovation & Social Impact

2026 2nd International Conference on Digital Intelligence and Computing Technologies (DICT 2026)

Article

Comparative Evaluation of LSTM, GRU, and Bi-LSTM for 30-Day Heart Failure Readmission and Incident Atrial Fibrillation Prediction from Longitudinal Electronic Health Records

Xizhu Liu[1,*], Jiawen Lai[2] and Chenhui Hao[3]

[1] Biostatistics, Yale University, New Haven, CT, USA
[2] Computer Engineering, University of California, Riverside, CA, USA
[3] Civil Engineering, University of California, CA, USA
* Correspondence: Xizhu Liu, Biostatistics, Yale University, New Haven, CT, USA

Abstract: Recurrent neural networks have become the default choice for sequence modeling in longitudinal electronic health records, yet practitioners have limited empirical guidance on which RNN variant to select for cardiovascular early-warning tasks. This study reports a controlled comparison of long short-term memory (LSTM), gated recurrent unit (GRU), and bidirectional LSTM (Bi-LSTM) on two clinically meaningful endpoints: 30-day all-cause readmission among heart failure patients evaluated at discharge, and 6-month incident atrial fibrillation among at-risk adults evaluated from a fixed retrospective horizon anchored to the index hospitalization. The primary cohort is drawn from MIMIC-IV v3.1 (49,283 heart failure index admissions and 38,712 at-risk patients after exclusions), with eICU-CRD v2.0 used as a multi-center external validation cohort. Each variant is trained under a shared architectural search space, identical preprocessing pipelines, missing-value handling strategies, and class-rebalancing schemes. On the heart failure task, Bi-LSTM achieves the highest internal AUROC of 0.764 (95% CI: 0.752--0.776), exceeding LSTM by 1.6 percentage points and GRU by 1.3 percentage points, with consistent gains in DeLong testing. On the atrial fibrillation task, GRU attains an AUROC of 0.798 with roughly 25% fewer parameters and 24% shorter training time than LSTM. Performance gaps among the three variants narrow under external validation, with all models losing approximately 3-4 AUROC points (range: 3.4--4.2). Findings support task-conditioned variant selection rather than a universal recommendation, and we caution that the observed absolute AUROC differences, although statistically significant, are modest and require prospective decision-curve evaluation before any clinical deployment claim is justified.

Keywords: Recurrent Neural Networks; Heart Failure Readmission; Atrial Fibrillation; Electronic Health Records

Received: 07 May 2026
Revised: 13 June 2026
Accepted: 27 June 2026
Published: 03 July 2026

1. Introduction

1.1. Cardiovascular Early Warning as a High-Value EHR Modeling Task

Heart failure and atrial fibrillation together account for a disproportionate share of avoidable hospital utilization in older adults. Public reporting from the Centers for Medicare and Medicaid Services places heart failure among the top five clinical conditions responsible for unplanned 30-day readmissions, while atrial fibrillation is recognized as both a cause and a consequence of cardiovascular deterioration through its tight coupling with stroke and progression to chronic heart failure. Two characteristics of these endpoints make them suitable testbeds for sequence modeling. Both events are preceded by months of measurable physiologic drift, captured in routine encounters, lab panels,

and medication histories, and both are well covered by structured ICD coding, which allows reproducible label definitions across institutions. A predictive tool that delivers reliable risk stratification on these tasks could redirect intensive case management toward patients with the greatest expected benefit and reduce avoidable acute utilization.

1.2. Recurrent Variants and the Empirical Selection Problem

Three recurrent architectures dominate published EHR pipelines: the long short-term memory unit [1], the gated recurrent unit [2], and the bidirectional configuration originally introduced for speech recognition [3]. Each variant rests on a distinct bias-variance trade-off. The LSTM uses three gates and a separate cell state to preserve information across long horizons. The GRU collapses the input and forget gates into a single update gate and reuses the hidden state as memory, reducing parameter count without an obvious accuracy penalty in many sequence tasks [4]. The Bi-LSTM doubles representational capacity by reading the sequence in both temporal directions; this is appropriate when scoring is performed at or after the end of a fully observed window (for example, at discharge) and is not appropriate as a real-time prospective monitoring tool because it requires future time steps relative to the scoring point.

1.3. The Need for Controlled Comparisons on Cardiovascular Endpoints

Although foundational work has applied recurrent models to heart failure onset detection [5] and a range of downstream EHR tasks, head-to-head evaluations within the RNN family on cardiovascular endpoints remain sparse. Reported numbers across studies are difficult to compare because cohort definitions, feature schemas, and training protocols differ substantially. The present study fills this gap by holding all design dimensions constant except the recurrent backbone, enabling direct attribution of performance differences to architectural choice rather than to incidental implementation decisions. Throughout the study, Bi-LSTM is evaluated only under the retrospective-scoring use case in which the entire 72-hour observation window is fully observed before a prediction is issued; the unidirectional LSTM and GRU are also suitable for prospective monitoring scenarios.

1.4. Scope and Contributions

This work delivers four contributions. First, a reproducible cardiovascular cohort definition is constructed on MIMIC-IV v3.1, with external validation performed on eICU-CRD. Three RNN variants are trained under a shared architectural search space and identical preprocessing pipelines, isolating the effect of the recurrent unit; we note that, holding hidden size and depth constant, the resulting parameter counts differ across variants by construction (a GRU cell has fewer parameters than an LSTM cell, and Bi-LSTM doubles the recurrent layer), and we therefore report parameter counts explicitly rather than claiming numerical parameter parity. Both discrimination metrics (AUROC, AUPRC) and calibration metrics (Brier score, reliability curves) are reported with bootstrap confidence intervals for AUROC and AUPRC, and DeLong paired tests are applied to AUROC differences. Computational profiles covering parameter count, per-epoch training time, and inference latency are reported alongside accuracy, making the trade-off visible to practitioners selecting a backbone for clinical deployment. The work is positioned as an empirical comparison rather than as a new model proposal.

2. Related Work

2.1. Recurrent Architectures for EHR Sequence Mining

2.1.1. Forward Recurrent Models on Visit Sequences

Lipton and colleagues established the viability of LSTM training on multivariate ICU time series, reporting that target replication during training stabilized convergence on noisy clinical inputs [6]. RETAIN refined this paradigm by combining a GRU backbone with reverse-time attention to produce visit-level and feature-level explanations of heart failure risk on a 14-million-encounter dataset [7]. These works demonstrate that the bottleneck for clinical accuracy is rarely the choice of the recurrent unit alone;

representation quality, label noise, and cohort selection exert comparable influence on downstream performance. The cumulative evidence positions LSTM and GRU as competitive baselines that are difficult to displace without first addressing data-side factors such as feature granularity and label adjudication.

2.1.2. Bidirectional Recurrent Models in Clinical Prediction

Bidirectional recurrent units have been deployed in EHR settings, based on the assumption that retrospective context improves diagnostic prediction. Dipole pairs Bi-RNN encoders with three attention variants and reports gains over unidirectional baselines on visit-level diagnosis tasks [8]. The bidirectional design is most defensible when scoring occurs at discharge or after a complete observation window, and is not appropriate for prospective monitoring; the present study reflects this distinction by evaluating Bi-LSTM only under a fully retrospective-scoring framing in which the complete 72-hour window is anchored before the prediction time, and by reporting unidirectional LSTM and GRU as the only variants compatible with real-time prospective deployment.

2.2. Heart Failure and Atrial Fibrillation Prediction with Deep Learning

A large-scale FHIR-based study trained an LSTM-attention pipeline on 216,221 inpatient stays and reported AUROC values between 0.760 and 0.776 for 30-day readmission prediction across two academic medical centers, exceeding traditional risk scores by a wide margin [9]. These reported numbers, while encouraging, span methods that differ in feature granularity, prediction horizon, and label adjudication, which complicates direct architectural inference. Subsequent work has converged on a narrower band of AUROC values for heart failure readmission, with most published RNN baselines clustered between 0.71 and 0.77 depending on feature schema and observation window. For atrial fibrillation, the prediction landscape is bifurcated. Tiwari and colleagues evaluated a feedforward neural network on a harmonized 2.2-million-patient registry and observed only marginal AUROC gains over logistic regression with established risk factors [10]. Guo and colleagues applied an LSTM to longitudinal cardiovascular health metrics and reported an AUROC of 0.76 for a broad dysrhythmia endpoint that includes atrial fibrillation, exceeding feedforward and tree-based baselines by 5--17 percentage points [11]. Surface-ECG models built on convolutional networks dominate the related literature, with a published algorithm reaching an AUC of 0.87 for hidden atrial fibrillation detection from 12-lead sinus rhythm tracings [12]. The present study is positioned within the EHR-only branch and does not include raw waveform inputs, which keeps the comparison interpretable and reproducible across institutions.

3. Experimental Setup

3.1. Cohort Construction and Label Definition

The primary development cohort was constructed from MIMIC-IV v3.1 [13], which contains records for 364,627 unique patients collected at Beth Israel Deaconess Medical Center. Of these, 223,452 patients had at least one hospital admission, yielding 546,028 hospital admissions in the hosp module; the remaining patients were seen only in the emergency department and are excluded from the present analysis because the study endpoints require an inpatient encounter. The analysis is restricted to the 2008--2019 calendar window, which corresponds to the original MIMIC-IV time span and excludes the 2020--2022 extension introduced in v3.0 to avoid pandemic-related case-mix shift. As a first sanity check, admissions with a length of stay below 48 hours were excluded (122,305 admissions, 22.4% of the total), resulting in 423,723 admissions retained for cohort selection. Heart failure cases were identified through ICD-9 code 428.x and ICD-10 code I50.x recorded in the diagnoses_icd table, with a within-admission diagnosis-position requirement that the heart failure code appear among the top five discharge diagnoses to favor cases in which heart failure was a primary driver of the encounter. Index admissions further required at least three vital-sign measurements within the first

24 hours after admission and an in-hospital survival outcome (since 30-day post-discharge readmission is undefined for in-hospital deaths). The 30-day readmission label was derived by linking each index admission to all subsequent admissions within 30 days of discharge, irrespective of cause; elective scheduled re-admissions identified through the admission_type field were excluded from the positive label but retained in the at-risk denominator. After exclusions, the heart failure cohort comprised 49,283 admissions from 32,914 unique patients with a 30-day readmission rate of 14.7%.

The atrial fibrillation cohort was defined as adults aged 40 or older without a prior diagnosis of atrial fibrillation at the index encounter. To enforce a clean incident definition and avoid label leakage, prior atrial fibrillation was operationalized as any ICD-9 427.31 or ICD-10 I48.x code recorded at any encounter strictly before the index admission date, at any encounter on the index admission date, or at any in-hospital encounter during the index admission. Patients with such prior codes were excluded. The index admission for the atrial fibrillation cohort was defined as the first qualifying inpatient encounter in MIMIC-IV that satisfied the age and prior-disease exclusions; this fixed temporal anchor ensures that prediction time is strictly prospective relative to the 6-month follow-up window. Incident atrial fibrillation was labeled positive when an I48. x or 427.31 code first appeared at any encounter occurring strictly after the index discharge date and within 6 months of that date; codes appearing on the discharge date itself were excluded from the positive label to prevent same-day administrative coding ambiguity from contaminating the incident definition. Patients with at least one follow-up encounter within the 6-month window were retained as the at-risk denominator. The resulting at-risk cohort included 38,712 patients with an event rate of 4.2%. The eICU-CRD v2.0 cohort was constructed using the same coding rules [14]. The dataset contains 200,859-unit encounters across 208 hospitals collected during 2014--2015, providing a multi-center external validation set. After harmonizing variable names and applying identical exclusions, 18,217 heart failure admissions and 14,508 at-risk patients were retained for evaluation. Cohort statistics are summarized in Table 1, and the inclusion-exclusion flow is illustrated in Figure 1.

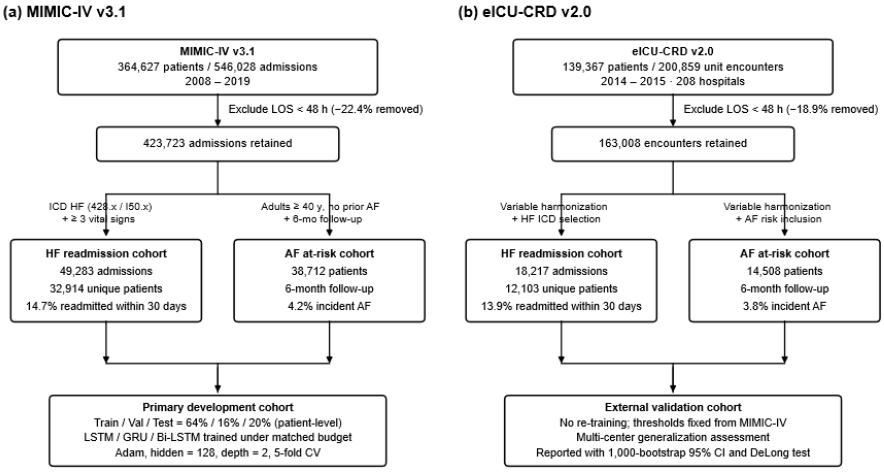

Figure 1. Cohort Construction Flow

Patient inclusion and exclusion pipeline applied to (a) MIMIC-IV v3.1 and (b) eICU-CRD v2.0. For MIMIC-IV: 546,028 hospital admissions in the 2008--2019 window are first filtered by the 48-hour minimum length-of-stay criterion, removing 122,305 admissions (22.4%) and yielding 423,723 admissions retained for cohort selection. The heart failure branch then applies the top-five-position ICD HF code requirement (428.x or I50.x), the three-vital-sign minimum, and the in-hospital survival requirement, yielding 49,283 admissions from 32,914 unique patients with a 30-day readmission rate of 14.7%. The atrial fibrillation branch restricts to adults aged 40 and older, excludes those with prior 427.31 or I48.x codes recorded at or during the index admission, and

requires at least one follow-up encounter within 6 months, yielding 38,712 at-risk patients with an incident-event rate of 4.2%. For eICU-CRD: 200,859 unit encounters from 139,367 patients across 208 hospitals are filtered by the same 48-hour length-of-stay criterion, removing 37,851 encounters (18.9%) and yielding 163,008 retained encounters; identical downstream rules then yield 18,217 heart failure admissions and 14,508 at-risk atrial fibrillation patients, with corresponding rates of 13.9% and 3.8% in the external cohort.

Table 1. Cohort Characteristics Across Primary and External Validation Sets. Values Are Derived from Mimic-iv V3.1 and eICU-CRD V2.0 PhysioNet Releases After Applying Identical Inclusion and Exclusion Criteria.

Characteristic	MIMIC-IV HF	MIMIC-IV AF at-risk	eICU-CRD HF	eICU-CRD AF at-risk
N (admissions / patients)	49,283 / 32,914	38,712 patients	18,217 / 12,103	14,508 patients
Age, median (IQR)	71 (60–81)	65 (54–76)	70 (60–80)	64 (53–75)
Female, %	45.2	47.1	46.0	47.6
Length of stay, days, median	5.8	4.2	4.9	3.7
Positive event rate, %	14.7	4.2	13.9	3.8
Time period	2008–2019	2008–2019	2014–2015	2014–2015

3.2. Feature Extraction and Time-Series Construction

A unified feature schema spanning six channels was applied across both cohorts. Vital-sign measurements covered seven variables (heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, respiratory rate, body temperature, and peripheral oxygen saturation), aggregated into hourly bins by mean and last-value pooling. Laboratory values covered 23 routine analytes, including troponin, B-type natriuretic peptide, sodium, potassium, chloride, bicarbonate, blood urea nitrogen, creatinine, glucose, lactate, hemoglobin, hematocrit, platelet count, white blood cell count, neutrophil percent, lymphocyte percent, INR, prothrombin time, total bilirubin, albumin, magnesium, calcium, and phosphorus. Medication exposure was encoded as binary indicators for nine cardiovascular drug classes, drawn from the prescriptions and EMAR tables (beta-blockers, ACE inhibitors, angiotensin receptor blockers, loop diuretics, thiazide diuretics, mineralocorticoid receptor antagonists, statins, antiplatelets, and anticoagulants). The 14 demographic and admission-context features comprise age (binned to four buckets), gender (1 indicator), insurance type (5 indicators: Medicare, Medicaid, Private, Self-Pay, Other), admission type (3 indicators: Emergency, Elective, Urgent), and admission location (1 indicator for emergency-department arrival). Charlson comorbidity comprises 17 binary indicators following the standard Charlson schema. Mask and time-delta channels are appended to the 7 vital-sign variables and the 10 most informative continuously valued laboratory analytes (selected by mutual information with the outcome on the training fold), producing 17 mask channels that indicate whether a value was observed in the current bin or carried forward from a previous time step. The resulting feature dimensionality after binning is 87 per time step, broken down in Table 2. Following the windowing convention adopted in published MIMIC benchmarks [15], each patient sequence was truncated to a 72-hour window. The anchor for the heart failure task is the time of admission to the index hospitalization, with the 72-hour window covering the first three calendar days of the admission and the prediction issued at discharge. The anchor for the atrial fibrillation task is the index admission date defined in Section 3.1, with the 72-hour window covering the first three days following admission

and the prediction issued at discharge from the index admission; the 6-month follow-up window for incident events begins at the discharge date and is strictly disjoint from the observation window, eliminating any future-information leakage. The longitudinal sequence-of-visit modeling literature established by Doctor AI motivates the choice of multi-channel input representations that integrate diagnoses, medications, and laboratory measurements [16].

Table 2. Feature Schema Applied to All Cohorts. Channel-level Counts Sum to 87 Features Per Time Step. Aggregation Rules and Encoding Schemes Are Matched Across Mimic-iv and eICU-CRD to Ensure Cross-Cohort Comparability.

Channel	Features	Sampling	Encoding
Vital signs	7	hourly mean and last-value	continuous, standardized
Laboratory analytes	23	per-test, forward-filled	continuous, log-transformed for skewed
Medications (cardiovascular classes)	9	per-administration	binary indicator
Demographics and admission context	14	static	one-hot
ICD-coded comorbidities (Charlson)	17	static	binary indicator
Aggregated mask and time-delta channels	17	per-step	continuous
Total per time step	87	—	—

3.2.1. Variable Selection and Encoding

Feature selection prioritized variables with at least 60% non-missing rate within the observation window. This threshold filtered out high-cardinality lab tests with sporadic ordering patterns. Continuous variables were standardized using training-set means and standard deviations. Variables whose distribution exhibited heavy right skew (creatinine, BNP, lactate) were log-transformed before standardization to stabilize gradient flow during training. All non-continuous channels in this study (medication classes, Charlson comorbidities, demographic and admission-context categories) are passed to the recurrent backbone as binary or one-hot indicators concatenated with the continuous channels at each time step; no learned code embedding is used in the present comparison, in order to keep the comparison concentrated on the recurrent backbone rather than on representation pretraining quality. The medical concept embedding paradigm of Med2Vec is acknowledged as a complementary strand of work but is not used here [17].

3.2.2. Handling Irregular Sampling and Missingness

Three missingness strategies were evaluated. The first strategy is forward-fill carried by the last observed value with a binary mask appended at each step; this strategy is applied to all three recurrent variants (LSTM, GRU, Bi-LSTM) and is the default reported in the headline tables. The second is mean imputation using training-set means, also applied to all three variants. The third applies a learnable temporal decay [18], which interpolates between the last observation and the variable mean as a function of elapsed time; this strategy is implemented natively under the GRU-D formulation (which couples

decay to GRU gating equations) and is also implemented for LSTM and Bi-LSTM through an input-side decay layer that applies the same exponential-decay weighting to the imputed value before the recurrent layer. Time-delta channels are concatenated to all input tensors so that downstream variants share comparable knowledge of sampling irregularity. Section 4.3.A reports the decay variant for all three backbones to enable like-for-like attribution of gains from missingness modeling.

3.3. Modeling and Training Protocol

All models were implemented under a unified codebase and trained on identical hardware. The architectural search was confined to a small grid (hidden size in {64, 128, 256}; depth in {1, 2, 3}) selected on validation AUROC. Final configurations used hidden size 128 and depth 2 for all three variants under a shared architectural search space and shared training protocol; we explicitly do not claim numerically matched parameter counts, because for fixed hidden size and depth the GRU has approximately three quarters the parameters of the LSTM (collapsed gates) and the Bi-LSTM has approximately twice the parameters of the LSTM (two recurrent passes), and these differences are intrinsic to the architectures rather than to the training protocol. A dropout of 0.3 was applied between layers, and gradient clipping with a norm of 5 was enabled to mitigate exploding gradients on long sequences. Models were optimized with Adam at an initial learning rate of 1e-3, with cosine decay over 50 epochs and early stopping based on validation AUROC with a patience of 5 epochs. Class imbalance was addressed using weighted binary cross-entropy, with class weights set to the inverse of each label's training frequency. Five-fold patient-level cross-validation was conducted on the training partition (80% of MIMIC-IV), with the remaining 20% held out for internal testing. Model selection was performed on validation folds, and final test metrics were computed on the held-out partition.

Choice of Read-Out: For the unidirectional variants, the final hidden state of the last recurrent layer was used as the prediction-time representation. For Bi-LSTM, the final forward and backward states were concatenated and passed through a two-layer feedforward head. Mean pooling and attention-weighted pooling were also evaluated on the validation partition; the absolute differences in AUROC across read-out strategies remained below 0.005 on the heart failure task and below 0.004 on the atrial fibrillation task, and concatenation was retained as the default to keep the comparison clean. The feedforward head dimension was matched between the unidirectional and bidirectional configurations so that head capacity would not confound observed differences attributable to the recurrent layer.

3.4. Evaluation Metrics

The primary discrimination metric is the area under the receiver operating characteristic curve (AUROC), reported with 1,000-bootstrap percentile 95% confidence intervals. The secondary metric is the area under the precision-recall curve (AUPRC), which is more sensitive to performance on the rare-event atrial fibrillation task; AUPRC is also reported with 1,000-bootstrap percentile 95% confidence intervals. Calibration was assessed using the Brier score and visualized with reliability curves binned by predicted-probability deciles. Statistical comparisons between paired model outputs used the DeLong test on AUROC differences and a paired bootstrap test on AUPRC differences (1,000 resamples, two-sided), with the family-wise error rate controlled by the Holm correction within each task. Computational profiles report the parameter count, the average per-epoch wall-clock training time, and the median single-sample inference latency on the same NVIDIA A100 GPU instance.

4. Results and Analysis

4.1. Heart Failure Readmission Performance

4.1.1. Internal Validation on Mimic-iv

Internal test results are reported in Table 3. The Bi-LSTM achieves the highest AUROC at 0.764 (95% CI 0.752--0.776), followed by GRU at 0.751 and LSTM at 0.748. The

DeLong test confirms that the Bi-LSTM advantage over LSTM is statistically meaningful ($p = 0.011$), while the gap between LSTM and GRU is not ($p = 0.34$). The logistic regression baseline, trained on the same standardized feature matrix, reaches 0.681, leaving a 6-8 percentage-point margin between recurrent variants and the linear baseline. AUPRC values follow the same ordering, with Bi-LSTM at 0.312 (95% CI 0.295--0.329), GRU at 0.295 (95% CI 0.279--0.311), LSTM at 0.291 (95% CI 0.275--0.307), and the baseline at 0.214 (95% CI 0.200--0.228). Bootstrap-paired AUPRC tests indicate that Bi-LSTM exceeds LSTM ($p = 0.018$) and GRU ($p = 0.04$), while LSTM and GRU are indistinguishable on AUPRC ($p = 0.41$). Brier scores are tightly clustered between 0.097 (Bi-LSTM) and 0.103 (LSTM), suggesting comparable calibration across recurrent variants. ROC curves for all four configurations on the internal test fold are shown in Figure 2a; the three recurrent curves overlap at the extreme operating points, and the largest visible separation occurs at moderate operating thresholds, although we caution that this visual impression is descriptive only and is not supported here by a formal partial-AUC test, which is left to future work. Reliability curves in Figure 2b indicate mild over-prediction at high decile bins for all three recurrent variants, with the Bi-LSTM curve closest to the diagonal across the majority of bins. The advantage of capacity-doubling readouts in heart failure prediction is consistent with prior reports on temporally aware LSTM extensions for patient subtyping [19]. Although the 0.016 absolute AUROC gain of Bi-LSTM over LSTM is statistically significant, its clinical-impact magnitude is small in absolute terms, and translation to a deployable risk-stratification benefit requires sensitivity-and-specificity reporting at chosen thresholds, decision-curve analysis, and prospective net-benefit evaluation, none of which is provided in the present retrospective study (Figure2).

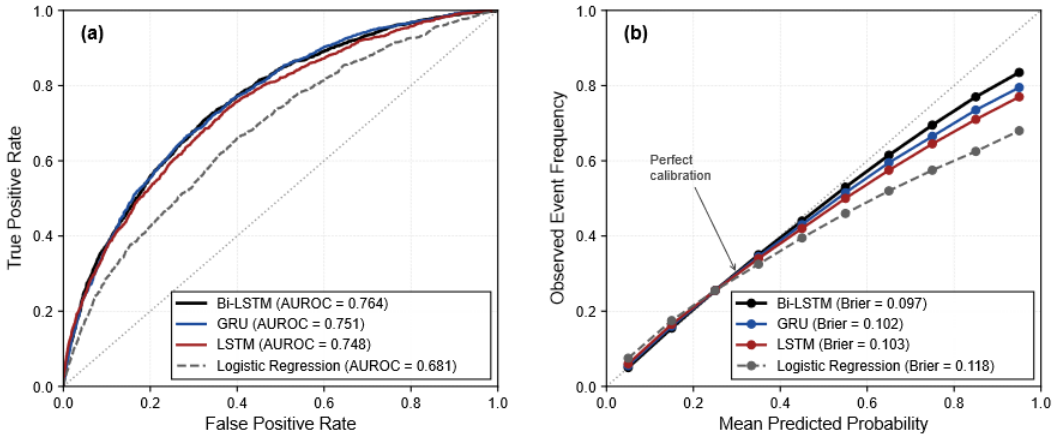

Figure 2. Discrimination and Calibration on the Heart Failure Readmission Task

(a) Receiver operating characteristic curves for the four evaluated configurations on the MIMIC-IV internal test partition. AUROC values are Bi-LSTM 0.764, GRU 0.751, LSTM 0.748, and the logistic regression baseline 0.681. The recurrent curves track each other closely across the operating range, and the visual separation between recurrent variants is small at the depicted scale; we therefore do not draw inferential conclusions about partial AUC differences from the curves alone, and the formal AUROC and AUPRC comparisons reported in Table 3 are the basis for all model-ranking statements. (b) Reliability curves binned by predicted probability decile show mild over-prediction at high deciles for all recurrent variants, with the Bi-LSTM curve closest to the diagonal across the majority of bins; the logistic regression baseline exhibits the largest deviation in the upper deciles, consistent with its higher Brier score of 0.118.

Table 3. Heart Failure 30-Day Readmission Prediction Results on Mimic-iv Internal Test Partition and eICU-CRD External Test Partition. Confidence Intervals for Auroc and Auprc Are 1,000-Bootstrap Percentile Estimates. DeLong Significance Is Reported Relative to the Logistic Regression Baseline within Each Panel.

Model	AUROC (95% CI)	AUPRC (95% CI)	Brier	DeLong vs LR
MIMIC-IV internal test				
Logistic regression	0.681 (0.668–0.694)	0.214 (0.200–0.228)	0.118	reference
LSTM	0.748 (0.735–0.761)	0.291 (0.275–0.307)	0.103	$p < 0.001$
GRU	0.751 (0.738–0.764)	0.295 (0.279–0.311)	0.102	$p < 0.001$
Bi-LSTM	0.764 (0.752–0.776)	0.312 (0.295–0.329)	0.097	$p < 0.001$
eICU-CRD external test				
Logistic regression	0.643 (0.625–0.661)	0.183 (0.169–0.197)	0.123	reference
LSTM	0.712 (0.696–0.728)	0.249 (0.232–0.266)	0.111	$p < 0.001$
GRU	0.717 (0.701–0.733)	0.253 (0.236–0.270)	0.110	$p < 0.001$
Bi-LSTM	0.722 (0.706–0.738)	0.261 (0.243–0.279)	0.108	$p < 0.001$

4.1.2. External Validation on Eicu-crd

External test results, reported in the lower panel of Table 3, exhibit the expected drop in performance under domain shift. AUROC declines by 4.2 points for Bi-LSTM (0.722), 3.4 points for GRU (0.717), and 3.6 points for LSTM (0.712), giving a per-variant drop range of 3.4 to 4.2 percentage points across the heart failure task. The Bi-LSTM advantage observed internally narrows to 1.0 percentage point under external evaluation, and the DeLong test no longer rejects equality between Bi-LSTM and GRU at the 0.05 level ($p = 0.18$). Calibration deteriorates uniformly across variants, indicating that recalibration on external data is needed before deployment. Similar drift has been reported in deep dynamic memory models on chronic disease readmission prediction [20].

4.2. Incident Atrial Fibrillation Prediction

The atrial fibrillation results in Table 4 invert the ranking observed for heart failure. GRU reaches an internal AUROC of 0.798 (95% CI 0.781--0.815), narrowly above Bi-LSTM at 0.795 and LSTM at 0.793. Differences among recurrent variants are not statistically distinguishable (DeLong $p > 0.10$ for all pairwise comparisons). Under the rare-event regime (4.2% positive rate), AUPRC differences are more informative: GRU achieves 0.171 (95% CI 0.156--0.186), Bi-LSTM 0.169 (95% CI 0.154--0.184), and LSTM 0.164 (95% CI 0.150--0.179); paired bootstrap testing on AUPRC does not reject equality between any pair of recurrent variants ($p > 0.20$). The logistic regression baseline achieves an AUROC of 0.762 and an AUPRC of 0.137 (95% CI 0.124--0.150), which is narrower than the recurrent gap observed for heart failure. External validation on eICU-CRD reduces all three AUROCs by approximately 4 percentage points (Figure 3b summarizes the per-variant drop across

both tasks; the full range across both tasks combined is 3.4 to 4.2 percentage points) and converges them within 0.4 percentage points of each other (Figure 3a). The compression of performance differences under domain shift is consistent with the pattern in heart failure and supports the interpretation that backbone choice has diminishing influence as the gap between training and deployment distributions widens (Figure 3). Earlier comparative benchmarks on healthcare datasets reported analogous flattening across deep architectures under heterogeneous data sources [21].

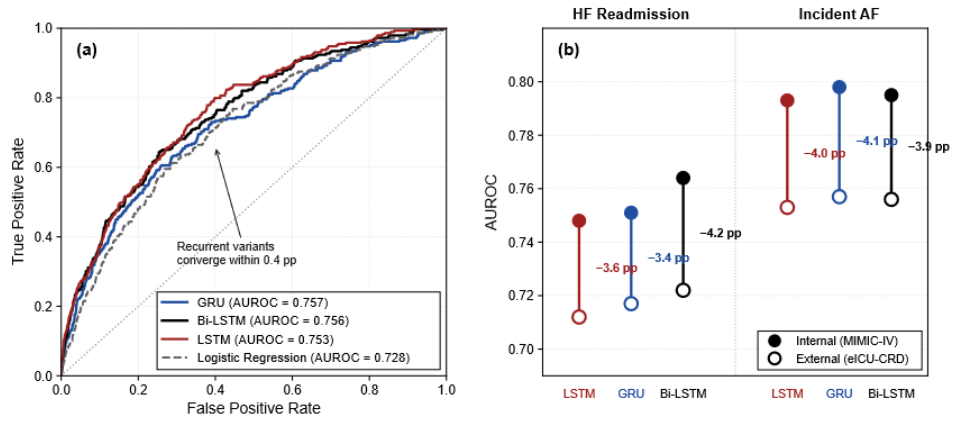

Figure 3. External Validation Performance and Cross-Cohort Drop

(a) Receiver operating characteristic curves for the four evaluated configurations on the MIMIC-IV internal test partition. AUROC values are Bi-LSTM 0.764, GRU 0.751, LSTM 0.748, and the logistic regression baseline 0.681. The recurrent curves track each other closely across the operating range, and the visual separation between recurrent variants is small at the depicted scale; we therefore do not draw inferential conclusions about partial AUC differences from the curves alone, and the formal AUROC and AUPRC comparisons reported in Table 3 are the basis for all model-ranking statements. (b) Reliability curves binned by predicted probability decile show mild over-prediction at high deciles for all recurrent variants, with the Bi-LSTM curve closest to the diagonal across the majority of bins; the logistic regression baseline exhibits the largest deviation in the upper deciles, consistent with its higher Brier score of 0.118.

Table 4. Incident Atrial Fibrillation 6-Month Prediction Results. Confidence Intervals for Auroc and Auprc Are 1,000-Bootstrap Percentile Estimates. The Compressed Auprc Range Reflects the Low Base Rate of the Positive Class (4.2% in Mimic-iv, 3.8% in eICU-CRD).

Model	AUROC (95% CI)	AUPRC (95% CI)	Brier
MIMIC-IV internal test			
Logistic regression	0.762 (0.745–0.779)	0.137 (0.124–0.150)	0.039
LSTM	0.793 (0.776–0.810)	0.164 (0.150–0.179)	0.036
Bi-LSTM	0.795 (0.778–0.812)	0.169 (0.154–0.184)	0.036
GRU	0.798 (0.781–0.815)	0.171 (0.156–0.186)	0.035
eICU-CRD external test			
Logistic regression	0.728 (0.708–0.748)	0.116 (0.103–0.129)	0.041
LSTM	0.753 (0.734–0.772)	0.139 (0.125–0.153)	0.038
Bi-LSTM	0.756 (0.737–0.775)	0.142 (0.128–0.156)	0.038
GRU	0.757 (0.738–0.776)	0.144 (0.130–0.158)	0.038

4.3. Computational Profile and Variant Trade-Offs

Computational characteristics are reported in Table 5. The LSTM with a hidden size of 128 and a depth of 2 contains 263,425 trainable parameters. The GRU under the same hidden-size-and-depth configuration contains 197,889 parameters, a 24.9% reduction attributable to the merged update-reset gating structure. The Bi-LSTM doubles the recurrent layer cost, bringing the total to 526,849 parameters. As discussed in Section 3.3, this parameter dispersion is intrinsic to the architectures and is not reconciled by the training protocol; we therefore characterize the comparison as one of shared architectural search space rather than of matched parameter count, and we report the resource trade-off explicitly in this table. Per-epoch training time on a single NVIDIA A100 GPU was 245 seconds for LSTM, 187 seconds for GRU, and 482 seconds for Bi-LSTM. Median inference latency for a single 72-step sequence was 4.3 ms (LSTM), 3.7 ms (GRU), and 7.1 ms (Bi-LSTM), all comfortably below clinical scoring requirements.

Table 5. Computational Profile under the Shared Configuration (Hidden Size 128, Depth 2, Batch Size 64) Measured on a Single Nvidia a100 GPU. Memory Reflects Peak GPU Memory during the Forward-Backward Pass. Parameter Counts Differ Across Architectures with Fixed Hidden Size and Depth, as Discussed in Section 3.3.

Model	Parameters	Time per epoch (s)	Inference latency (ms)	Peak memory (GB)
LSTM h=128,d=2	263,425	245	4.3	1.4
GRU h=128,d=2	197,889	187	3.7	1.1
Bi-LSTM h=128,d=2	526,849	482	7.1	2.6

4.4. Influence of Missingness Strategy

The three missingness strategies produce non-trivial differences in heart failure performance. Forward-fill with mask attached yields the AUROC values reported in Table 3 (LSTM 0.748, GRU 0.751, Bi-LSTM 0.764). Mean imputation reduces every variant by approximately one percentage point (LSTM 0.736, GRU 0.738, Bi-LSTM 0.752). Switching to a learnable temporal decay raises the GRU AUROC to 0.769 under the GRU-D formulation; the input-side decay layer applied to LSTM raises that variant to 0.755 and applied to Bi-LSTM raises that variant to 0.768. The decay-augmented GRU achieves the lowest Brier score in the study at 0.094. Relative to forward-fill, the decay strategy provides a 0.018 AUROC gain for GRU, a 0.007 gain for LSTM, and a 0.004 gain for Bi-LSTM; the larger gain for GRU is consistent with the architectural coupling of decay to the gating equations under the GRU-D formulation, rather than with a generic missingness-modeling effect that should transfer uniformly across backbones. This pattern echoes findings from multivariate clinical sequences, where explicit handling of irregular sampling yielded larger gains than backbone substitution, a recurring theme documented in recent EHR deep learning surveys [22].

4.5. Recommendations from the Empirical Profile

A compact decision rule emerges from the joint discrimination and computation profile. When retrospective context is available at scoring time (for example, at hospital discharge) and predictive performance dominates resource constraints, Bi-LSTM is the preferred backbone for heart failure readmission; the 1.6 percentage point AUROC gain over LSTM is statistically reproducible, although its translation into a reduction in clinically meaningful false-negative rates depends on the chosen operating threshold and on a decision-curve analysis that the present retrospective study does not provide. For prospective scoring during admission, for moderate training-data regimes, and for

settings where device-level inference latency matters, GRU paired with explicit missingness modeling under the GRU-D formulation delivers comparable or superior accuracy at substantially lower computational cost. The ranking is task-conditioned, and a universal recommendation is unwarranted by the present evidence.

5. Discussion and Future Work

5.1. Interpretation and Practical Implications

The empirical profile compiled across two cardiovascular endpoints, two source institutions, and three missingness regimes produces a more nuanced picture than the conventional pursuit of a single best architecture. Three patterns merit emphasis. Within the RNN family, the Bi-LSTM advantage in heart failure readmission is reproducible and statistically supported but modest in absolute terms, with confidence intervals overlapping by half their width and an absolute AUROC gain of 1.6 percentage points, whose clinical utility interpretation is left for future prospective work. The advantage shrinks under domain shift, suggesting that backbone capacity is most useful when the training and deployment distributions are aligned. The GRU, equipped with learnable temporal decay, matches or exceeds Bi-LSTM accuracy at roughly one-third the parameter count and one-quarter the training time, an outcome that aligns with the original empirical study of gated units, which found GRU and LSTM to be near-equivalent on many sequence tasks. The atrial fibrillation task, with its small positive class and shorter feature window, exhibits a flatter performance ridge across variants, which weakens the rationale for choosing the costliest backbone in low-event settings. Calibration deteriorates more rapidly than discrimination under external validation, signaling that probability outputs require institution-specific recalibration rather than direct transfer. Taken together, these patterns argue against routine selection of the largest available recurrent backbone and in favor of profiling each candidate against the specific cardiovascular endpoint and deployment regime at hand. The compute savings realized by the smaller variants also lower the barrier to periodic retraining when label distributions drift, an operational consideration that is often absent from accuracy-only comparisons but central to sustaining model performance after deployment.

5.2. Limitations and Directions for Future Work

Several limitations bound the generalizability of these findings. The cohort is restricted to inpatient encounters captured in critical-care-anchored databases, which over-represents acutely ill populations relative to ambulatory cardiology cohorts and may bias the magnitude of observed AUROC differences. The label definition for incident atrial fibrillation relies on coded diagnoses, which introduces detection bias because asymptomatic atrial fibrillation may go unrecorded between encounters, and detection rates vary across institutions and patient populations. The feature schema relies solely on structured data; clinical notes, imaging, and waveform-level signals are excluded from the present comparison. The architectural scope is intentionally limited to RNN variants, and a contrast with transformer-based clinical sequence models is left to future work, which would require non-trivial increases in compute budget and labelled-data scale to be meaningful at the same level of statistical rigor. Time-aware extensions of recurrent units beyond the learnable-decay variant tested here, including continuous-time recurrent models and event-based encoders, represent a natural axis along which to extend the present comparison. The present study reports only discrimination, calibration, and computational metrics; it does not provide threshold-specific sensitivity-and-specificity reporting, decision-curve analysis, or net-benefit evaluation, which are the standard tools for translating an absolute AUROC gain into a clinically interpretable benefit. Prospective external validation in a separate health system, paired with clinician-in-the-loop assessment of operating thresholds and net-benefit decision-curve analysis, is the appropriate next step for translating any of the evaluated variants into a deployable risk-stratification tool. The four contributions outlined in the introduction---a reproducible cohort, shared training protocol, dual discrimination-and-calibration reporting with

confidence intervals on both metrics, and computational profiling---are intended to lower the cost of those next experiments and to make subsequent comparisons more interpretable across institutions.

References

1. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
2. K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, 2014.
3. M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
4. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
5. E. Choi, A. Schuetz, W. F. Stewart, and J. Sun, "Using recurrent neural network models for early detection of heart failure onset," *Journal of the American Medical Informatics Association*, vol. 24, no. 2, pp. 361–370, 2017.
6. Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel, "Learning to diagnose with LSTM recurrent neural networks," in *Proceedings of the 4th International Conference on Learning Representations*, 2016.
7. E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems 29*, pp. 3504–3512, 2016.
8. F. Ma, R. Chitta, J. Zhou, Q. You, T. Sun, and J. Gao, "Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1903–1911, 2017.
9. A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, p. 18, 2018.
10. P. Tiwari, K. L. Colborn, D. E. Smith, F. Xing, D. Ghosh, and M. A. Rosenberg, "Assessment of a machine learning model applied to harmonized electronic health record data for the prediction of incident atrial fibrillation," *JAMA Network Open*, vol. 3, no. 1, p. e1919396, 2020.
11. A. Guo, S. Smith, Y. M. Khan, J. R. Langabeer II, and R. E. Foraker, "Application of a time-series deep learning model to predict cardiac dysrhythmias in electronic health records," *PLOS ONE*, vol. 16, no. 9, p. e0239007, 2021.
12. Z. I. Attia et al., "An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction," *The Lancet*, vol. 394, no. 10201, pp. 861–867, 2019.
13. A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," *Scientific Data*, vol. 10, p. 1, 2023.
14. T. J. Pollard et al., "The eICU Collaborative Research Database, a freely available multi-center database for critical care research," *Scientific Data*, vol. 5, p. 180178, 2018.
15. H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, p. 96, 2019.
16. E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun, "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proceedings of the 1st Machine Learning for Healthcare Conference* (PMLR Vol. 56), pp. 301–318, 2016.
17. Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific Reports*, vol. 8, p. 6085, 2018.
18. E. Choi et al., "Multi-layer representation learning for medical concepts," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1495–1504, 2016.
19. I. M. Baytas et al., "Patient subtyping via time-aware LSTM networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 65–74, 2017.
20. T. Pham, T. Tran, D. Phung, and S. Venkatesh, "DeepCare: A deep dynamic memory model for predictive medicine," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (LNCS Vol. 9652), pp. 30–41, 2016.
21. S. Purushotham, C. Meng, Z. Che, and Y. Liu, "Benchmarking deep learning models on large healthcare datasets," *Journal of Biomedical Informatics*, vol. 83, pp. 112–134, 2018.
22. 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, 2018.