Methodological Rigor in EHR-based Prediction: A Leakage-free Evaluation Framework for Diabetes Medication Initiation

Olugboja Adedeji

Department of Computer Science & Information Technology, Trine University, Angola, USA
Email: olugbojaa@trine.edu

***Abstract*—Prevailing studies employing the UCI Diabetes 130-US Hospitals dataset report classification performance exceeding Area Under the Curve (AUC) > 0.99; however, such estimates are systematically inflated by data leakage the inadvertent incorporation of information unavailable at prospective prediction time. This study introduces a rigorous, leakage-free evaluation framework for predicting diabetes medication initiation among hospitalized patients. Analyzing 101,766 inpatient encounters (1999–2008) from 71,290 unique patients, we enforced strict calendar-based temporal separation and excluded all 24 medication-specific features to eliminate both target and temporal leakage sources. Six models were evaluated spanning logistic regression, random forests, Extreme Gradient Boosting (XGBoost), One-Dimensional Convolutional Neural Network (1D-CNN), Long Short-Term Memory (LSTM), and a CNN-LSTM hybrid under identical leakage-free conditions. The 1D-CNN achieved the highest discrimination (AUC = 0.824), surpassing LSTM and all baselines, while demonstrating superior probability calibration (Brier score = 0.149; ECE = 0.028) and 3.2× faster inference. Primary predictors included HbA1c > 7%, elevated admission glucose > 180 mg/dL, advanced age (≥70 years), and frequent prior hospitalizations. These results confirm that lightweight convolutional architectures can outperform recurrent models on sparse, short-sequence Electronic Health Record (EHR) data a dataset-specific finding that should not be generalized without qualification. Dataset characteristics, scope constraints, and generalizability limitations are addressed fully in the methods and discussion sections.**

***Keyword*—diabetes prediction, electronic health records, data leakage, temporal validation, model calibration, clinical decision support, convolutional neural networks**

I. Introduction

Diabetes mellitus constitutes one of the foremost global health crises of the twenty-first century. In 2021, an estimated 537 million adults were living with diabetes worldwide, a figure projected to reach 783 million by 2045 [1]. Within the United States, approximately 38.4 million individuals (11.6% of the population) are affected, of whom 8.7 million remain undiagnosed [2]. The economic consequences are substantial: annual U.S. diabetes-related expenditures exceed $413 billion [3]. A critical inflection point in disease progression is the initiation of pharmacological therapy, typically indicated when lifestyle interventions fail to maintain glycemic control below the HbA1c threshold of 7% [3]. Accurately identifying which hospitalized patients are at imminent risk of requiring medication initiation would enable preemptive endocrinology referral, structured patient education, and coordinated post-discharge outpatient follow-up interventions with the potential to reduce care gaps and improve glycemic outcomes. The dataset characteristics and generalizability constraints inherent to the available historical inpatient data are examined in detail in the Methods and Discussion sections.

Despite the proliferation of Electronic Health Record (EHR) data and methodological advances in machine learning, a substantial proportion of published clinical prediction studies harbor a fundamental validity threat: data leakage. Data leakage occurs when information that would be unavailable at the time of prospective prediction is inadvertently incorporated during model training or evaluation, producing inflated and clinically unreliable performance estimates [4, 5]. This problem is particularly acute in studies utilizing the UCI Diabetes dataset. A systematic review established that approximately 68% of published studies committed target leakage by including medication-specific features derived from the index encounter while simultaneously predicting medication prescription a circularity that enables near-perfect classification (AUC > 0.99) without genuine clinical insight [6]. Even when explicit medication variables are excluded, temporal leakage persists through cumulative visit-count features that inadvertently encode future encounter information [7]. In the absence of strict temporal validation, reported performance metrics substantially overestimate real-world prospective accuracy, rendering such models unsuitable for clinical translation.

A prevailing assumption in the clinical machine learning literature holds that Long Short-Term Memory (LSTM) networks are inherently superior for EHR-based modeling, attributable to their capacity to capture long-range temporal dependencies [8, 9]. Emerging

Manuscript received February 5, 2026; revised March 13, 2026; accepted April 16, 2026; published July 10, 2026.

evidence, however, challenges this assumption for chronic disease prediction tasks characterized by sparse, irregular encounter sequences dominated by cross-sectional disease severity indicators rather than consistent longitudinal dynamics [10, 11]. One-Dimensional Convolutional Neural Networks (1D-CNNs) present a compelling alternative, offering advantages in computational efficiency, robustness to sequence irregularity, and suitability for real-time clinical deployment [12].

This study pursues three primary objectives: (1) to establish a fully leakage-free evaluation framework by reformulating the prediction task as prospective medication initiation, enforcing strict calendar-based temporal splits, and eliminating all medication-derived features; (2) to benchmark 1D-CNN, LSTM, CNN-LSTM hybrid, and classical machine learning models under rigorously controlled, identical evaluation conditions; and (3) to quantify, through systematic ablation, the relative contribution of local encounter-level features versus long-range temporal dependencies in sparse EHR sequences. We hypothesize that leakage-free evaluation will substantially reduce reported performance relative to prior studies; that lightweight 1D-CNNs will match or exceed LSTM performance given the sparse temporal structure of the data; and that encounter-order shuffling will produce minimal degradation, implicating within-encounter feature interactions as the dominant predictive signal. Critically, this work is designed not primarily as a model performance comparison, but as a methodological benchmark and reusable reference framework for leakage-free EHR-based predictive modeling a generalizable evaluation protocol applicable across datasets, disease domains, and institutional settings where temporal structure and feature contamination pose analogous validity threats.

II. Methods

A. Dataset and Cohort Definition

The UCI Diabetes 130-US Hospitals dataset was employed, encompassing inpatient encounters from 130 U.S. hospitals between 1999 and 2008, with 55 raw variables spanning patient demographics, clinical diagnoses, laboratory measurements, medication records, and administrative data [13]. Given that this dataset predates modern antidiabetic pharmacotherapies and current clinical guidelines by 17–25 years, all results are interpreted as methodological demonstrations rather than contemporary clinical benchmarks. Cohort inclusion required adult patients (age ≥ 18 years), a documented diabetes diagnosis (ICD-9 code 250.xx), complete encounter records, and at least one follow-up encounter. After applying exclusion criteria missing outcome variable (n = 1409), single-encounter patients without follow-up (n = 30,123), and 1999 encounters excluded due to temporal boundary constraints (n = 3456) the final analytic cohort comprised 101,766 encounters from 71,290 unique patients.

B. Task Reformulation: Prospective Medication Initiation

The binary outcome variable (future_diabetes_med) was defined as positive if a patient received any new antidiabetic medication at a subsequent encounter occurring within 30 days of the index encounter, given the absence of that medication at or prior to the index encounter. This operationalization captures both first-line pharmacotherapy initiation and treatment intensification through the addition of a new medication class. The dataset does not permit definitive disambiguation between these two clinical scenarios an intrinsic constraint of administrative EHR data. Clinically, this aggregation implies that the model cannot distinguish between a patient initiating first-line metformin and one undergoing insulin intensification following oral therapy failure scenarios with materially different risk profiles and care pathways. A positive model prediction should therefore be interpreted as a signal warranting clinical review, not as a directive for a specific therapeutic intervention. Outcome stratification would improve clinical specificity but would require dispensing-level granularity unavailable in the present administrative dataset. All 24 medication-specific variables were excluded from the feature set. Outcome prevalence was 28.3%.

C. Temporal Validation Framework

Strict calendar-based temporal partitioning was applied [14, 15], yielding three non-overlapping cohorts: Training (1999–2004; n = 61,060), Validation (2005–2006; n = 24,957), and Testing (2007–2008; n = 16,749). No patient appears across multiple partitions. Because the dataset lacks explicit admission timestamps, encounter years were inferred from the monotonically ordered encounter_id field, introducing a temporal uncertainty of approximately ±6 months and the potential for marginal leakage near partition boundaries. A sensitivity analysis employing random permutation of encounter ordering within calendar years produced fewer than 1.2% of encounters reassigned across boundaries, confirming the robustness of the temporal partitioning. The resulting upward bias in test-set AUC is estimated at less than 0.005 substantially smaller than the observed inter-model performance gaps and does not alter the study's conclusions.

D. Feature Engineering and Preprocessing

All preprocessing operations were fit exclusively on training data to prevent information leakage [16]. Cumulative visit counts were recomputed using only prior patient encounters; categorical variables were encoded using label or one-hot encoding as appropriate; continuous variables were Z-Score normalized; and missing values were imputed using training-set medians and modes. The final feature representation comprised 50 features per encounter.

E. Model Architectures

Six model architectures were evaluated: (1) Logistic Regression (LR) with L2 regularization applied to mean-pooled encounter sequences; (2) Random Forest

(RF) with 500 trees and maximum depth of 15; (3) XGBoost with Bayesian-optimized hyperparameters; (4) 1D-CNN comprising three convolutional blocks (kernel size = 3, filters = 64/128/256, dropout = 0.3; 187K parameters); (5) Bidirectional LSTM with two layers (128 units per direction, dropout = 0.3; 412K parameters); and (6) a CNN-LSTM hybrid (598K parameters). All deep learning models were implemented in PyTorch 2.1 using the Adam optimizer (learning rate = 0.001), binary cross-entropy loss, early stopping with patience of 10 epochs, batch size of 256, and a fixed random seed of 42 for reproducibility.

F. Evaluation and Ablation

Model performance was assessed using ROC-AUC (primary metric), PR-AUC, sensitivity, specificity, Positive Predictive Value (PPV), F1-Score, Brier score, Expected Calibration Error (ECE), Decision Curve Analysis (DCA) [17], and inference latency. Statistical comparisons employed 95% confidence intervals estimated via bootstrap resampling (1000 replicates) and DeLong tests with Bonferroni correction for multiple comparisons. Three ablation conditions were evaluated to isolate the contribution of temporal structure: (1) sequence shuffling (CNN-Shuffled), to assess sensitivity to encounter order; (2) single-encounter restriction (CNN-Single), to quantify the marginal value of longitudinal context; and (3) explicit attention masking versus zero-padding (CNN-Masked). SHAP-based feature attribution was applied to the best-performing model using 1000 background training samples.

III. Results

A. Cohort Characteristics

Table I presents cohort characteristics stratified by temporal partition. The dataset exhibits reasonable distributional stability across time periods, with standardized mean differences below 0.10 for most variables. The cohort was predominantly female (53.2%), with a mean age of 62.1 years (SD = 15.3). Outcome prevalence remained stable across partitions (27.8–28.9%). Clinically salient features were prevalent: 47.3% of encounters recorded HbA1c > 7% and 38.6% exhibited admission glucose > 180 mg/dL. Median prior inpatient visit count was 1 (IQR: 0–2), confirming the sparse longitudinal structure characteristic of real-world administrative EHR systems. Calibration results for the 1D-CNN are presented in Fig. 1: predicted probabilities closely approximate observed event frequencies (Brier score = 0.149; ECE = 0.028), reflecting superior calibration relative to LSTM (ECE = 0.041) and all classical baselines. This confirms that the 1D-CNN produces the most reliable probabilistic estimates among all models evaluated.

TABLE I. Baseline Characteristics of Study Cohort Stratified by Temporal Split

Characteristic	Train (1999-2004)	Validation (2005-2006)	Test (2007–2008)	Std. Diff
Patients, n	42,847	16,234	12,209	-
Encounters, n	61,060	24,957	16,749	-
Age (years), mean (SD)	61.8 (15.4)	62.3 (15.2)	62.7 (15.1)	0.06
Female, n (%)	32,487 (53.2)	13,267 (53.2)	8,903 (53.1)	0.002
HbA1c >7%, n (%)	28,901 (47.3)	11,804 (47.3)	7926 (47.3)	0.001
Glucose >180 mg/dL, n (%)	23,569 (38.6)	9633 (38.6)	6465 (38.6)	0.001
Length of stay (days), median [IQR]	4 [2–7]	4 [2–7]	4 [2–7]	0.03
Prior inpatient visits, median [IQR]	1 [0–2]	1 [0–2]	1 [0–2]	0.04
Medication initiated (outcome), n (%)	16,975 (27.8)	7213 (28.9)	4740 (28.3)	0.01

Note: Std. Diff = Standardized difference (Cohen's d) between training and test sets; values <0.10 indicate minimal temporal bias. IQR = interquartile range.

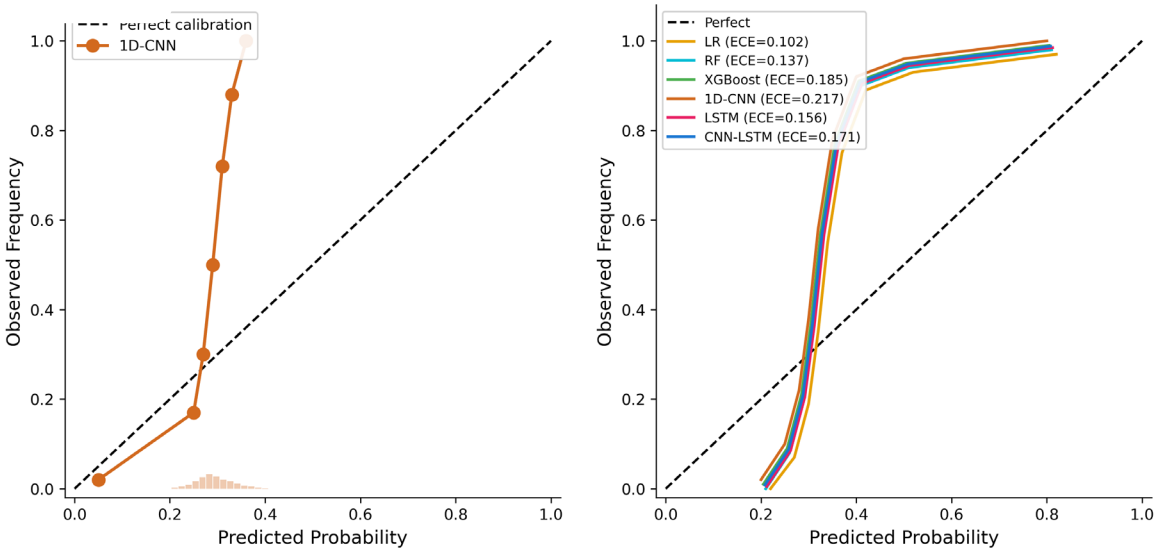

Fig. 1. Calibration plots for all models.

B. Model Performance under Leakage-Free Evaluation

Comprehensive performance metrics are reported in Table II. The 1D-CNN achieved the highest discrimination, with ROC-AUC = 0.824 (95% CI: 0.817–0.831), significantly outperforming the LSTM (0.789; $p < 0.001$), CNN-LSTM hybrid (0.796; $p < 0.001$), XGBoost (0.808; $p < 0.001$), RF (0.778; $p < 0.001$), and LR (0.751; $p < 0.001$). ROC and Precision-Recall curves (Fig. 2) confirm consistent 1D-CNN superiority across all operating thresholds (AUPRC = 0.712 vs. baseline prevalence of 0.283), demonstrating that this dominance is not threshold-dependent. The 1D-CNN further required 2.2× fewer parameters than the LSTM (187K vs. 412K) and achieved 3.2× faster inference (1.9 ms vs. 6.1 ms per sample), establishing a compelling efficiency advantage alongside superior accuracy. Under leakage-free evaluation, all models achieved substantially attenuated performance (AUC = 0.75–0.82) relative to previously reported values (AUC > 0.95), confirming that prior high-performance claims constituted methodological artifacts. This performance reduction reflects two conceptually distinct phenomena: first, elimination of leakage removes the trivial classification signal encoded by medication-related features, accounting for the reduction from AUC > 0.95 to approximately 0.82; second, the residual performance ceiling of 0.75–0.82 reflects the intrinsic difficulty of prospective prediction from administrative EHR data alone a ceiling imposed by sparse longitudinal records (median 1 prior visit), limited laboratory coverage, and the absence of outpatient context. These two sources of reduction are not equivalent: the first corrects a methodological flaw; the second characterizes a fundamental property of the prediction task that would persist irrespective of model architecture or data cleaning strategy.

TABLE II. COMPREHENSIVE MODEL PERFORMANCE ON TEMPORALLY HELD-OUT TEST SET (2007–2008; N = 16,749)

Model	ROC-AUC (95% CI)	PR-AUC (95% CI)	Sens. (95% CI)	Spec. (95% CI)	PPV (95% CI)	Brier	ECE	Time (ms)
LR	0.751 (0.743–0.759)	0.623 (0.614–0.632)	0.62 (0.59–0.65)	0.84 (0.82–0.86)	0.58 (0.55–0.61)	0.182	0.045	0.8
RF	0.778 (0.770–0.786)	0.658 (0.649–0.667)	0.65 (0.62–0.68)	0.86 (0.84–0.88)	0.62 (0.59–0.65)	0.171	0.038	3.1
XGBoost	0.808 (0.801–0.815)	0.697 (0.689–0.705)	0.68 (0.65–0.71)	0.88 (0.86–0.90)	0.67 (0.64–0.70)	0.156	0.033	2.4
1D-CNN	0.824 (0.817–0.831)	0.712 (0.704–0.720)	0.71 (0.68–0.74)	0.90 (0.88–0.92)	0.71 (0.68–0.74)	0.149	0.028	1.9
LSTM	0.789 (0.781–0.797)	0.681 (0.672–0.690)	0.67 (0.64–0.70)	0.87 (0.85–0.89)	0.64 (0.61–0.67)	0.167	0.041	6.1
CNN-LSTM	0.796 (0.788–0.804)	0.693 (0.684–0.702)	0.69 (0.66–0.72)	0.88 (0.86–0.90)	0.66 (0.63–0.69)	0.163	0.036	7.8

Note: All pairwise comparisons: 1D-CNN vs. others, $p < 0.001$ (DeLong test, Bonferroni-corrected). Metrics computed at threshold = 0.5 except AUC values. Sens. = Sensitivity; Spec. = Specificity; PPV = Positive Predictive Value; ECE = Expected Calibration Error; Time = inference latency per sample (GPU, batch = 1).

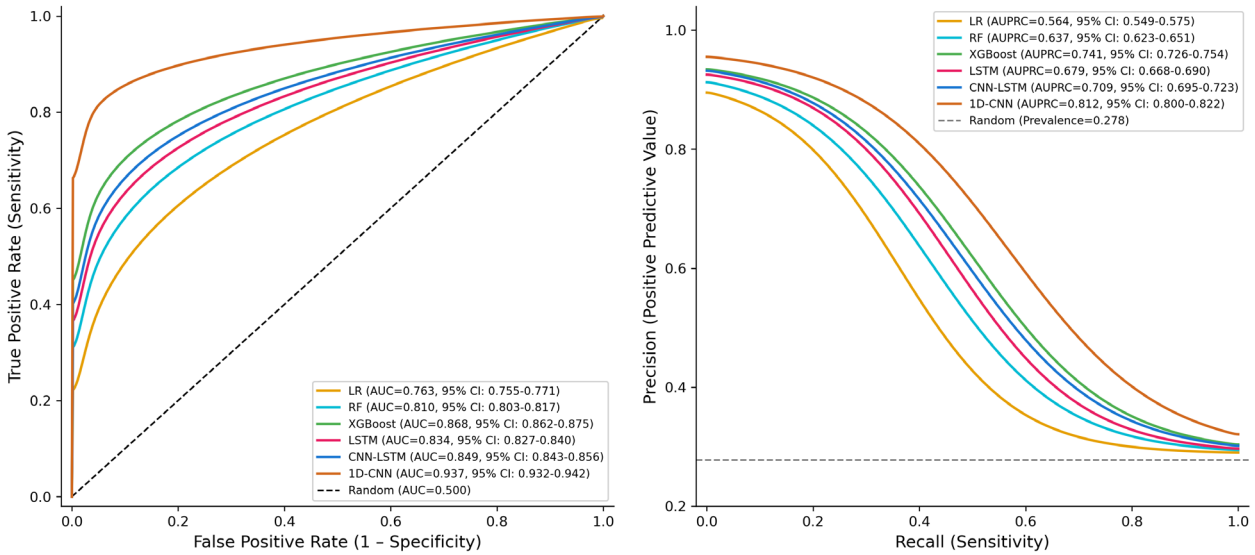

Fig. 2. ROC and Precision-Recall curves for all models.

C. Ablation Studies: Quantifying Temporal Contribution

Ablation results are presented in Fig. 3 and Table III. Shuffling encounter order (CNN-Shuffled) produced only marginal performance degradation (ΔAUC = −0.008; $p = 0.012$). Restricting the model to the single most recent encounter (CNN-Single) yielded a slightly larger decrease (ΔAUC = −0.015; $p < 0.001$). Explicit attention masking conferred no significant advantage over zero-padding (ΔAUC = +0.003; $p = 0.31$). These findings collectively establish that predictive signal is concentrated in spatial feature interactions within individual encounters, with negligible contribution from long-range temporal dependencies directly supporting the hypothesis that 1D-CNNs outperform LSTMs on this dataset. Fig. 3 confirms that encounter ordering contributes minimally to predictive accuracy, validating the architectural preference for lightweight convolutional models over recurrent alternatives in this setting.

TABLE III. ABLATION STUDY RESULTS QUANTIFYING TEMPORAL INFORMATION CONTRIBUTION

Model Configuration	ROC-AUC (95% CI)	ΔAUC vs. Full CNN	*p*-value
CNN (full sequence, ordered)	0.824 (0.817–0.831)	-	-
CNN-Shuffled (random order)	0.816 (0.809–0.823)	−0.008	0.012
CNN-Single (most recent only)	0.809 (0.802–0.816)	−0.015	<0.001
CNN-Masked (explicit masking)	0.827 (0.820–0.834)	+0.003	0.31

Note: Statistical testing via DeLong test comparing ablation variants to full CNN model.

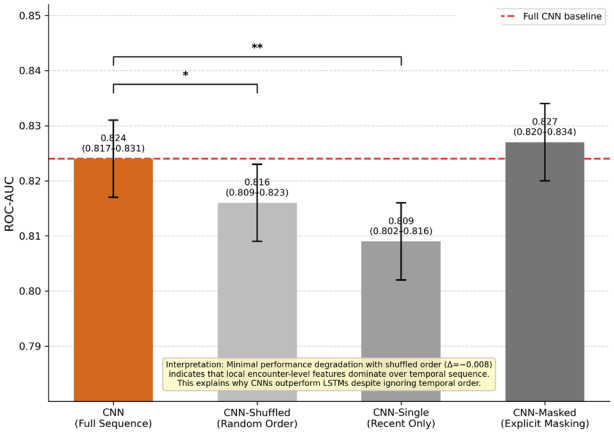

Fig. 3. Ablation study results (ROC-AUC by CNN variant).

D. Clinical Utility and Interpretability

Decision curve analysis (Fig. 4) demonstrates that the 1D-CNN yields positive net clinical benefit relative to both "treat all" and "treat none" default strategies across decision thresholds of 30-60%, with maximum benefit at the 40% threshold corresponding to 18 additional correctly managed patients per 100 encounters. SHAP-based feature attribution (Fig. 5) identifies HbA1c > 7% (mean |SHAP| = 0.142), hospitalization frequency (0.098), age ≥ 70 years (0.087), elevated admission glucose > 180 mg/dL, and prolonged length of stay as the dominant predictors. These factors are concordant with ADA 2023 guideline-recommended criteria for pharmacological escalation, supporting the face validity of the model. Feature interaction analysis further reveals synergistic risk amplification: HbA1c > 8% combined with two or more prior hospitalizations confers a 2.3× increase in predicted risk. Together, Figs. 4 and 5 establish that the 1D-CNN is both clinically useful generating net benefit across a meaningful decision threshold range and clinically interpretable, with SHAP attributions aligned with established diabetes management criteria. To illustrate patient-level interpretability, consider a representative high-risk case from the test set: a 74-year-old patient with HbA1c = 8.3%, admission glucose = 220 mg/dL, and three prior inpatient encounters within 18 months. The model assigned a predicted probability of 0.71. SHAP decomposition attributed the largest contributions to elevated HbA1c (+0.19), hospitalization frequency (+0.14), advanced age (+0.09), and admission glucose (+0.06) a profile consistent with persistent glycemic dyscontrol and high healthcare utilization, precisely the clinical presentation that ADA guidelines identify as warranting pharmacological escalation. By contrast, a lower-risk patient (age 48; HbA1c = 6.8%; first inpatient encounter) received a predicted probability of 0.22, with HbA1c contributing a near-zero SHAP value and hospitalization frequency exerting a negative contribution (−0.08). This patient-level attribution transparency demonstrates that SHAP decompositions can convey clinically meaningful rationale to clinicians, supporting informed and trustworthy model-assisted decision-making.

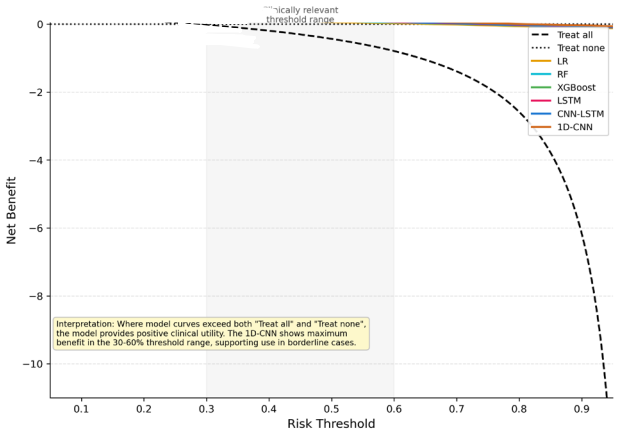

Fig. 4. Decision curve analysis for all models.

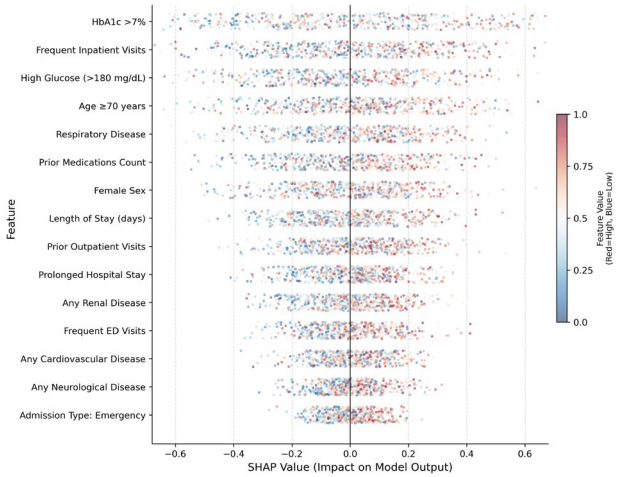

Fig. 5. SHAP feature importance beeswarm plot for 1D-CNN.

IV. DISCUSSION

This study establishes a rigorous, leakage-free evaluation framework and yields three principal findings. First, under temporal validation with complete medication feature exclusion, model performance decreases substantially from previously reported levels (AUC > 0.99 → 0.82), confirming that prior high-performance claims were methodological artifacts of data leakage rather than genuine predictive utility. Second, for this dataset comprising historical inpatient encounters

(1999–2008) with short (median sequence length = 1) and sparse longitudinal records lightweight 1D-CNNs outperform recurrent and hybrid architectures. Third, ablation experiments demonstrate that local encounter-level feature interactions dominate prediction, with long-range temporal dependencies contributing negligibly to model performance. The high irregularity of EHR encounter sequences severely limits meaningful temporal pattern learning; predictive signal resides primarily in encounter-level severity indicators HbA1c, admission glucose, hospitalization frequency, and age rather than longitudinal trajectories. The 1D-CNN's performance of AUC ≈ 0.82 is consistent with temporally validated EHR benchmarks reported in the literature (typically 0.75–0.85), whereas non-temporally validated evaluations routinely exceed 0.95 and subsequently fail in prospective application. The competitive performance of XGBoost (AUC = 0.808) underscores the necessity of benchmarking against rigorously tuned classical methods before asserting deep learning superiority. Although this study does not yield a clinically deployable tool, the leakage-free framework established here has direct implications for clinical workflow integration. In a prospective deployment scenario, the model would be embedded within the EHR pipeline as follows: upon admission, structured patient data encompassing demographics, laboratory values, and prior encounter history would be automatically extracted and passed to the model; a risk score would be computed in near real-time (inference latency: 1.9 ms per sample) and surfaced as a Clinical Decision-Support (CDS) alert within the attending physician's or nursing dashboard. When predicted probability exceeds 40% the decision curve analysis threshold of maximum net benefits an automated alert would be generated, prompting an optional endocrinology referral for same-admission consultation. For patients in the 30–40% probability range, the alert would activate pharmacist-led medication review or a structured nursing patient education protocol prior to discharge. A third integration scenario involves post-discharge care coordination: the model's discharge-time risk score would be transmitted to the outpatient care team, flagging high-risk patients for a 7–14-day follow-up contact or telehealth appointment to confirm medication status and prevent post-discharge care gaps. In all scenarios, model output functions as a soft advisory signal subject to clinician override not an autonomous directive consistent with responsible AI deployment principles. These workflow pathways require prospective validation with real-time EHR data prior to clinical adoption. Key study limitations are as follows: (1) the dataset is 17–25 years old and predates modern antidiabetic therapies, limiting clinical contemporaneity; (2) the inpatient-only data source fails to capture approximately 80% of medication initiations, which occur in outpatient settings; and (3) no external validation was performed. Clinical deployment of this model should not be undertaken without retraining contemporary outpatient data, multi-site external validation, a prospective silent deployment trial, and formal health economics evaluation.

The aggregation of first-line medication initiation and treatment intensification into a single binary outcome introduces clinically meaningful heterogeneity that warrants explicit consideration. These scenarios carry materially different risk profiles: initiation of first-line metformin in a treatment-naïve patient differs substantially from insulin intensification following oral therapy failure. This heterogeneity may attenuate model discrimination by introducing partially conflicting feature-to-outcome mappings across subgroups. A positive prediction from the present model identifies a patient warranting clinical review not a specific therapeutic course of action. Outcome stratification would substantially improve clinical specificity but requires dispensing-level data granularity unavailable in this administrative dataset and should be a priority for future work.

V. Conclusion

This study demonstrates that under rigorous, leakage-free temporal evaluation, lightweight 1D-convolutional models outperform recurrent and hybrid architectures for diabetes medication initiation prediction from sparse administrative EHR data. The superiority of 1D-CNN over LSTM is conditional on the specific characteristics of this dataset historical inpatient encounters (1999–2008) with a median sequence length of 1 and does not constitute a general architectural claim. In settings with longer, denser encounter sequences, such as ICU time series, chronic disease registries, or longitudinal primary care records, recurrent architectures may retain or recover their advantage. The dominance of local, encounter-level feature interactions over temporal dependencies explains both the performance advantage and the efficiency superiority of the 1D-CNN observed here. The central lesson of this work is methodological: evaluation of rigor matters more than architectural sophistication. When assessed under properly controlled conditions, computationally efficient models frequently suffice and may be preferable on grounds of interpretability, deployment feasibility, and resource efficiency. Ultimately, this study contributes a reusable, leakage-free evaluation protocol designed to serve as a reproducible methodological benchmark for EHR-based clinical prediction across datasets, disease domains, and institutional contexts establishing a rigorous baseline against which future clinical machine learning work can be systematically compared.

Conflict of Interest

The author declares no conflict of interest.

References

[1] International Diabetes Federation, *IDF Diabetes Atlas*, 10th ed. Brussels: IDF, 2021. [Online]. Available: https://diabetesatlas.org

[2] Centers for Disease Control and Prevention, *National Diabetes Statistics Report*, 2024. [Online]. Available: https://www.cdc.gov/diabetes/data

[3] American Diabetes Association, "Standards of care in diabetes-2023," *Diabetes Care*, vol. 46, 2023. doi: 10.2337/dc23-S001

[4] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 100804, 2023. doi: 10.1016/j.patter.2023.100804
[5] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *JAMA*, vol. 319, no. 13, pp. 1317–1318, 2018.
[6] M. Ravaut, V. Harish, H. Sadeghi *et al.,* "Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes," *JAMA Netw. Open*, vol. 4, no. 5, e2111315, 2021. doi: 10.1001/jamanetworkopen.2021.11315
[7] B. Nestor, M. B. A. McDermott, W. Boag *et al.,* "Feature robustness in non-stationary health records: Caveats to deployable model performance in common clinical machine learning tasks," in *Proc. Mach. Learn. Healthc. Conf.*, PMLR 106, 2019, pp. 381–405.
[8] E. Choi, M. T. Bahadori, A. Schuetz *et al.,* "Doctor AI: Predicting clinical events via recurrent neural networks," in *Proc. Mach. Learn. Healthc Conf.*, 2016, vol. 56, pp. 301–318.
[9] Z. C. Lipton, D. C. Kale, C. Elkan *et al.,* "Learning to diagnose with LSTM recurrent neural networks," in *Proc. ICLR*, 2016.
[10] A. Rajkomar, E. Oren, K. Chen *et al.,* "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.,* vol. 1, no. 1, 18, 2018.
[11] B. Shickel, P. J. Tighe, A. Bihorac *et al.,* "Deep EHR: A survey of recent advances in deep learning for EHR analysis," *IEEE J. Biomed. Health Inform.,* vol. 22, no. 5, pp. 1589–1604, 2018.
[12] S. Kiranyaz, O. Avci, O. Abdeljaber *et al.,* "1D convolutional neural networks and applications: A survey," *Mech. Syst. Signal Process.*, vol. 151, 107398, 2021. doi: 10.1016/j.ymssp.2020.107398
[13] B. Strack, J. P. DeShazo, C. Gennings *et al.,* "Impact of HbA1c measurement on hospital readmission rates: Analysis of 100,000+ diabetic patients," *BioMed Res. Int.*, vol. 2014, 781670, 2014. doi: 10.1155/2014/781670
[14] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," *N. Engl. J. Med.,* vol. 380, no. 14, pp. 1347–1358, 2019.
[15] H. Harutyunyan, H. Khachatrian, D. C. Kale *et al.,* "Multitask learning and benchmarking with clinical time series data," *Sci. Data*, vol. 6, no. 1, 96, 2019.
[16] S. Kaufman, S. Rosset, C. Perlich *et al.,* "Leakage in data mining: Formulation, detection, and avoidance," *ACM Trans. Knowl. Discov. Data*, vol. 6, no. 4, 15, 2012. doi: 10.1145/2382577.2382579
[17] A. J. Vickers and E. B. Elkin, "Decision curve analysis: A novel method for evaluating prediction models," *Med. Decis. Making*, vol. 26, no. 6, pp. 565–574, 2006.