International Journal of Computer Engineering in Research Trends
ISSN: 2349-7084/ https://doi.org/10.22362/ijcert/2024/v11/i12/v11i1205
Volume 11, Issue 12, December 2024

Research Paper

Silent Model Degradation in Clinical AI Detecting and Quantifying Undocumented Data Drift in Live EHR Systems

[1*] Sreeja Poduri, [2] Chavali Sri Gowri, [3] Lavanya Addepalli, [4] Jaime Lloret

[1] Independent Researcher, Salt Lake City, Utah, USA*
Email ID: sreejap1997@gmail.com
[2] Chandigarh University, Kharar, Punjab
Email ID: chavaligowri95@gmail.com
[3] Department of Communication and Cultural Industries, Universitat Politècnica de València, Spain,
Email ID: phani.lav@gmail.com, ORCID: 0000-0002-2651-163X
[4] Instituto de Investigación para la Gestión Integrada de Zonas Costeras,
Universitat Politècnica de València, Valencia, Spain,
Email ID: jlloret@dcom.upv.es

**Corresponding Author(s): sreejap1997@gmail.com*

Received: 14/09/2024 Revised: 18/10/2024 Accepted: 20/12/2024 Published: 31/12/2024

Abstract: Clinical prediction models implemented within clinical healthcare are getting high stakes clinical decision support, however, they are seldom subjected to post deployment performance other than simple validation. This paper draws attention to and defines silent model degradation a failure mode where predictive validity deteriorates over time as a result of undocumented data drift failing to initiate standard system warnings. We will use a large longitudinal electronic health record (EHR) dataset in a tertiary care hospital system to show that both covariate drift and concept drift increase gradually over time following deployment and result in more gradual changes in discrimination, calibration, and prediction stability. To overcome this risk, we design CLIOPS, a combined framework of post-deployment monitoring that combines temporal drift, longitudinal performance deterioration modelling, and unsupervised early-warnings without depending on instant outcome classification. On comparative analysis, CLIOPS is less prone to operational load and detects degradation sooner and more steadily than current feature-based and label-dependent drift detection procedures. These results demonstrate that accurate performance loss leading to clinical consequences may be hidden and that safe and reliable implementation of clinical AI must be accompanied with label-free monitoring.

Keywords: Clinical artificial intelligence**;** Silent model degradation**;** Data drift detection**;** post-deployment monitoring**;** Electronic health records**;** AI governance and safety

---

1. Introduction

Machine learning-based clinical prediction models find more and more application as part of electronic health record (EHR) systems to aid high-stakes clinical and operational decisions. Though these models usually perform highly in development and post hoc validation, they do not usually undergo reliability evaluation after implementation and so their reliability is usually assumed. This is a very dangerous assumption when used in healthcare environments where patient groups, clinical activities and documentation procedures change over time, and cause changes in the method of data generation that are often informal, or written down. One of the main lessons of this article is that in clinical settings, model failure may not be an extreme or noticeable moment. Predictive performance can instead degrade over time due to the buildup of undocumented covariate and concept drift resulting in what we call silent model degradation. This failure mode leads to clinically significant discrimination,

calibration, and prediction stability declines without recommended system alerts or governance. Due to the nature of clinical workflows and outcome labels, which are usually delayed and not monitored post-deployment, such degradation may continue to go unnoticed, which is likely to subject patients and health care systems to long-term stakes [1-3].

The second important lesson is that currently, post-deployment monitoring strategies lack the ability to identify this phenomenon individually. The perceived downstream performance impairment of feature-level statistical drift tests is usually insensitive, whereas the use of label-dependent change detectors is limited by lagging outcomes. Besides, the majority of monitoring strategies consider the issues of data drift, prediction behavior, and performance evaluation individually, without reflecting on the combined time dynamics and accumulative impact. Because of this, the existing monitoring practices are more likely to spot the degradation when a significant loss in performance has taken place. In order to overcome these shortcomings, a comprehensive, longitudinal surveillance model of deployed clinical artificial intelligence systems is suggested in this paper. The framework incorporates a temporal decomposition of live EHR data and the observance of undocumented covariate and concept drift, the explicit modeling of time-varying performance degradation, and the self-report early-warning signals without acquiring outcome labels at any point in time. The proposed method combined with treating deployment as a continuous process and not a definite outcome helps to identify silent degradation earlier and reliably besides reducing the burden of operational requirements [4-6].

The following are the outlines of this paper. Part 2 presents a literature review of the previous research on data drift, concept drift, and post-deployment monitoring in clinical machine learning and identifies limitations of the current solutions. Section 3 outlines the research methodology involving the formal formulation of the problem, mechanism of drift identification, and automatic early-warning policies. In section 4, the empirical findings provided based on a large longitudinal EHR dataset prove the popularity of silent model degradation and the efficiency of the suggested framework. It is then followed by the implications of the results on clinical safety, operational governance, and regulatory oversight as well as limitations of the study, as discussed in section 5. Lastly, Section 6 wraps up the paper, and gives the future research directions such as multimodal data extension, multi-institutional validation, and adaptive remediation policies to deployed clinical AI systems.

2. Literature Review

The use of machine learning models for clinical prediction has expanded rapidly across a wide range of healthcare applications, including risk stratification, early warning systems, and operational decision support. Numerous studies have demonstrated that models trained on electronic health record (EHR) data can achieve high predictive performance under retrospective evaluation. However, most published work evaluates models at development time, often using random or temporally split test sets, with limited attention to model behavior after deployment. As a result, post-deployment performance is frequently assumed to be stable rather than empirically monitored. Recent work has begun to challenge this assumption by documenting performance degradation of clinical models over time. These studies highlight that healthcare environments are inherently dynamic, shaped by evolving clinical guidelines, patient demographics, care delivery processes, and documentation practices. Nevertheless, much of the existing literature focuses on describing degradation after it has occurred, rather than developing systematic mechanisms for its early detection under realistic operational constraints [7-9].

2.1 Data Drift in Electronic Health Records

Data drift refers to changes in the statistical properties of input data over time and is commonly categorized as covariate drift, label drift, or concept drift. In EHR-based modeling, covariate drift may arise from shifts in patient populations, measurement practices, or coding behaviors, while concept drift reflects changes in the relationship between predictors and outcomes due to evolving clinical interventions or standards of care. Several studies have shown that EHR data are particularly susceptible to drift, even in the absence of explicit system changes. Feature distributions may shift gradually as clinical workflows evolve, new diagnostic tests are introduced, or care pathways are modified. Importantly, these changes often occur without formal documentation, making them difficult to detect using conventional system monitoring. While the presence of drift in EHR data is now well recognized, its downstream impact on deployed model performance remains underexplored in longitudinal, real-world settings [10-11].

2.2 Drift Detection Methods

A broad range of statistical and algorithmic methods have been proposed to detect data drift. Feature-level statistical tests, such as population stability index (PSI), Kolmogorov–Smirnov tests, and distributional divergence measures, are widely used due to their simplicity and label-free nature. However, these methods operate

independently on individual features and do not directly assess whether detected drift has meaningful implications for model predictions or clinical outcomes. As a result, they are prone to false alerts driven by benign fluctuations and may fail to detect harmful shifts that manifest only through complex feature interactions. Label-dependent drift detection methods, including adaptive windowing and error-rate monitoring approaches, aim to detect changes in model performance directly. While these methods can be sensitive to degradation, they rely on timely access to outcome labels, which is often infeasible in clinical settings where outcomes may be delayed by hours, days, or longer. This dependency limits their practical utility for early warning in real-world healthcare deployments. More recently, prediction-based and uncertainty-based monitoring approaches have been explored, leveraging changes in model confidence, calibration, or prediction entropy as proxies for degradation. These methods offer promise as label-efficient alternatives but are typically evaluated in isolation and lack integration with explicit data drift analysis or longitudinal performance modelling [12-13].

2.3 Post-Deployment Monitoring and Model Governance

Post-deployment monitoring of clinical AI systems has increasingly been recognized as a governance and safety requirement, particularly in the context of regulatory guidance for AI-based medical software. Emerging frameworks emphasize the need for continuous performance evaluation, transparency, and post-market surveillance. However, operational implementations of these principles remain limited, and monitoring practices are often reduced to periodic audits or threshold-based alerts. Critically, most existing monitoring approaches do not account for the cumulative and time-dependent nature of model degradation. By treating drift detection and performance evaluation as pointwise or episodic tasks, these approaches fail to capture gradual degradation trajectories and their compounding effects. Moreover, few frameworks explicitly address the trade-off between early detection and operational burden, a key consideration in clinical environments where excessive alerts can undermine trust and usability [14-15].

2.4 Research Gap

Despite growing recognition of data drift and post-deployment model failure in clinical machine learning, several critical gaps remain. First, there is a lack of empirical frameworks that characterize silent model degradation as a longitudinal phenomenon driven by undocumented drift, rather than as isolated performance drops or abrupt failures. Second, existing drift detection methods are rarely evaluated in terms of their ability to provide early, operationally feasible warnings under realistic constraints such as delayed outcome labels and fixed deployed models. Third, current approaches tend to treat data drift, prediction behavior, and performance decay as separate problems, without integrating them into a unified monitoring strategy. This work addresses these gaps by proposing a comprehensive, longitudinal monitoring framework that integrates covariate and concept drift detection, temporal performance modeling, and unsupervised early-warning signals into a single governance-oriented approach. By empirically evaluating this framework on real-world EHR data, the study advances the understanding of silent model degradation and provides practical tools for sustaining the safety and reliability of deployed clinical AI systems.

3. Proposed Methodology

The section introduces a common methodological approach to identifying and measuring the silent model decadence of operational clinical artificial intelligence systems on live electronic health record (EHR) information. The suggested solution is meant to work within realistic post-deployment constraints, such as the unavailability of outcomes in time, non-parametric models, and unregistered changes in the system. The treatment of deployment as a longitudinal process and not an assessment point helps the methodology to assess the data stability, predictability and performance integrity continually over time. The framework combines temporal breaks of post-deployment data streams with statistical detection of covariate and concept drift, explicit modelling of time-dependent performance decays and unsupervised early-warning mechanisms, which are not based on outcome labels at inference times. The combination of these elements offers a theory-based and practically viable policy of observing deployed clinical models in active healthcare settings that assists in active governance and risk elimination during a model lifecycle.

3.1 Problem Definition

Let $\mathcal{D}_0 = \{(x_i, y_i)\}_{i=1}^{N}$ denote the dataset used to train a clinical prediction model $f_\theta: \mathcal{X} \to \mathcal{Y}$, where $x_i \in \mathbb{R}^d$ represents a vector of structured EHR features (e.g., vitals, labs, demographics), $y_i \in \{0,1\}$ denotes a clinical outcome, and θ are learned model parameters. The model is trained under the assumption that future data are drawn from the same joint distribution $P_0(X,Y)$. After deployment at time t_0, the model is exposed to a sequence of temporally ordered data streams $\{\mathcal{D}_t\}_{t=t_1}^{T}$, where each $\mathcal{D}_t \sim P_t(X,Y)$. In real clinical environments, the data-generating process evolves such that $P_t(X,Y) \neq P_0(X,Y)$,

even in the absence of explicit system changes. We define silent model degradation as the progressive loss of predictive validity of f_θ caused by latent shifts in P_t that are neither documented nor detected by healthcare IT monitoring systems. The methodological objective of this study is to (i) detect undocumented drift in EHR data streams, (ii) quantify its contribution to model performance decay, and (iii) identify early-warning signals of impending prediction failure without relying on labeled outcomes at inference time.

Figure 1 provides a summary of the methodology that is suggested in order to identify silent model degradation in already deployed clinical artificial intelligence systems. The framework shows such assurance of the data contained within post-deployment electronic health records over time is not only partitioned and statistically examined to determine undocumented covariate and concept drift, but it also measures time-dependent performance decays, and then produces unmonitored early-warning signals. The novel philosophy of combining drift detection with longitudinal performance monitoring allows evaluating the model integrity of live clinical environments continuously, without explicit system modifications or post-hoc audits

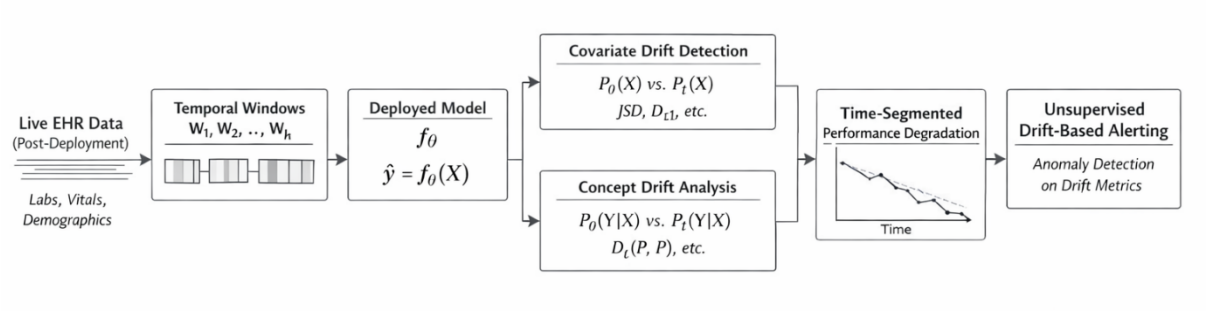


Fig.1. Proposed Architecture

3.2 Temporal Segmentation of Live EHR Data

To enable longitudinal analysis, the post-deployment data stream is partitioned into non-overlapping temporal windows $\{\mathcal{W}_k\}_{k=1}^{K}$, where each window corresponds to a fixed interval (e.g., monthly or quarterly). Each window $\mathcal{W}_k$ contains feature observations $X_k = \{x_{k,1}, \dots, x_{k,n_k}\}$ and, when available, delayed outcome labels Y_k. This segmentation allows the empirical estimation of time-indexed marginal and conditional distributions $\hat{P}_k(X)$ and $\hat{P}_k(Y \mid X)$, enabling drift analysis while preserving the temporal ordering inherent to clinical workflows.

3.3 Covariate Drift Detection via Distributional Divergence

Undocumented covariate drift is assessed by comparing feature distributions across temporal windows. For each feature $j \in \{1, \dots, d\}$, we estimate its marginal distribution $\hat{P}_k(X^{(j)})$ within window k. Drift magnitude is quantified using a symmetric divergence metric, such as the Jensen-Shannon divergence:

$$\mathrm{JSD}\big(\hat{P}_0^{(j)} \| \hat{P}_k^{(j)}\big) = \frac{1}{2}\mathrm{KL}\big(\hat{P}_0^{(j)} \| M^{(j)}\big) + \frac{1}{2}\mathrm{KL}\big(\hat{P}_k^{(j)} \| M^{(j)}\big)$$

where $M^{(j)} = \frac{1}{2}\big(\hat{P}_0^{(j)} + \hat{P}_k^{(j)}\big)$ and $\mathrm{KL}(\cdot)$ denotes the Kullback-Leibler divergence.

Aggregating across features yields a window-level drift score:

$$D_k^{\mathrm{cov}} = \frac{1}{d}\sum_{j=1}^{d} \mathrm{JSD}\big(\hat{P}_0^{(j)} \| \hat{P}_k^{(j)}\big)$$

This formulation allows identification of features exhibiting high latent drift despite unchanged schemas, providing insight into undocumented shifts in clinical documentation or practice.

3.4 Concept Drift and Prediction-Outcome Decoupling

While covariate drift captures changes in $P(X)$, silent degradation may also arise from concept drift, where the conditional relationship $P(Y \mid X)$ evolves. To quantify this effect, we evaluate the stability of the prediction-outcome relationship over time.

Let $\hat{y}_{k,i} = f_\theta(x_{k,i})$ be the model prediction for instance i in window k. Concept drift is measured by examining temporal changes in conditional outcome likelihoods:

$$\Delta_k^{\mathrm{concept}} = \left|\mathbb{E}[Y_k \mid \hat{y}_k \in \mathcal{B}] - \mathbb{E}[Y_0 \mid \hat{y}_0 \in \mathcal{B}]\right|$$

for prediction bins $\mathcal{B} \subset [0,1]$. This bin-conditional analysis reveals whether identical model confidence scores

correspond to different clinical risks over time, a hallmark of silent failure in deployed AI systems.

3.5 Time-Segmented Performance Decay Modeling

To quantify degradation explicitly, model performance metrics (e.g., AUROC, Brier score) are computed within each temporal window:

$$\mathcal{M}_k = \mathcal{L}\left(Y_k, \hat{Y}_k\right)$$

where $\mathcal{L}(\cdot)$ denotes a suitable loss or evaluation function. Performance decay is modeled as a function of elapsed deployment time τ_k :

$$\mathcal{M}_k = \beta_0 + \beta_1 \tau_k + \epsilon_k$$

A statistically significant negative slope β_1 indicates systematic post-deployment degradation. Crucially, this regression is evaluated alongside drift metrics D_k^{cov} and $\Delta_k^{\text{concept}}$ to establish causal associations between undocumented drift and performance loss.

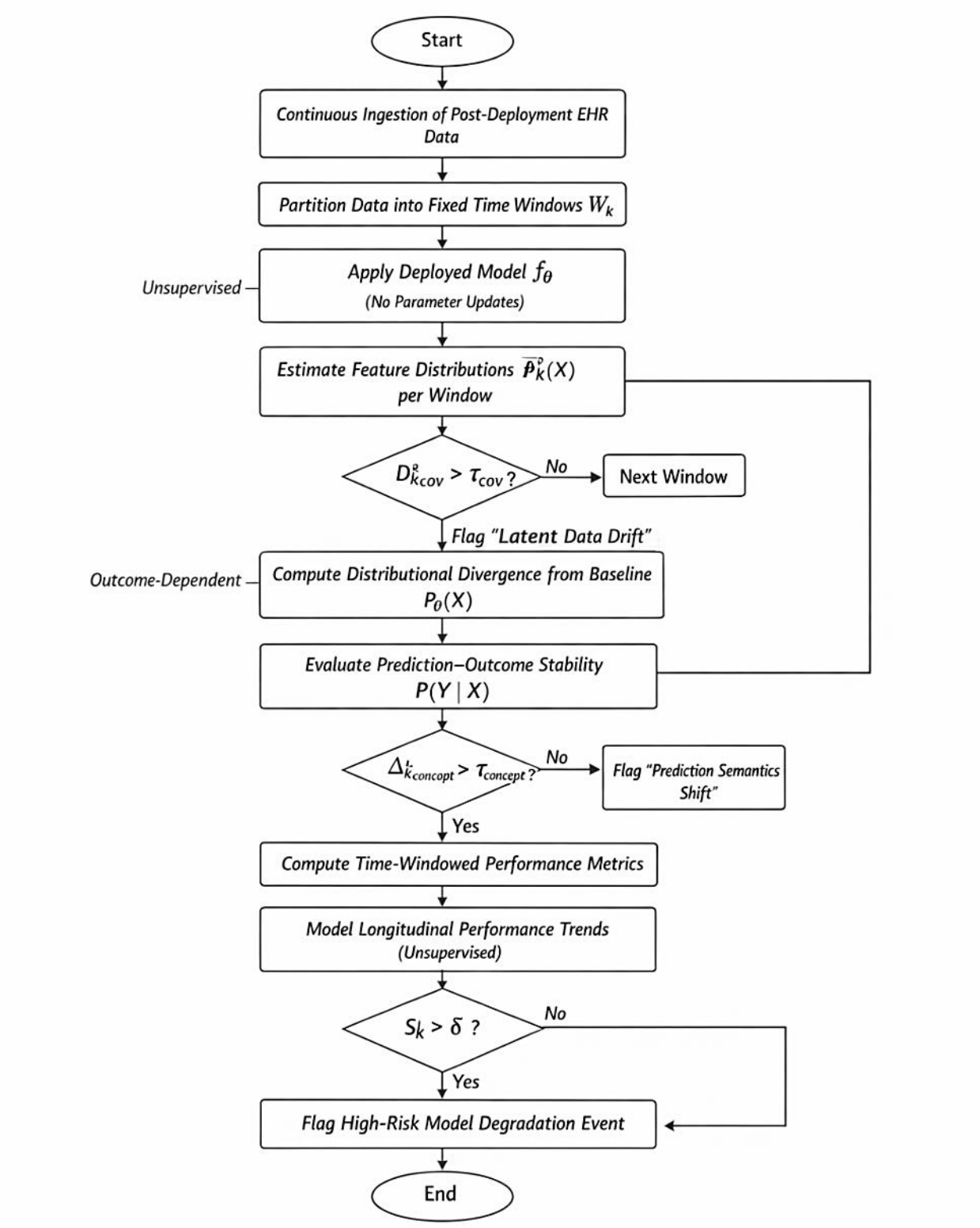
3.6 Unsupervised Early-Warning Detection

To enable proactive governance, we propose an unsupervised alerting mechanism that flags windows exhibiting anomalous drift patterns before outcome labels are available. Let $\mathbf{z}_k = [D_k^{\text{cov}}, \text{Var}(\hat{y}_k), H(\hat{y}_k)]$, where $H(\cdot)$ denotes prediction entropy. An anomaly score is computed via density estimation:

$$S_k = -\log p(\mathbf{z}_k)$$

where $p(\cdot)$ is estimated from early post-deployment windows assumed to be stable. Windows exceeding a predefined threshold δ are flagged as high-risk for silent degradation.

Figure 2 shows the implementation plan of the suggested framework of monitoring deployed clinical AI systems. The flowchart outlines the sequential processing visuals needed in order to ingest live EHR information, do a temporal segmentation, make model predictions, recognize undocumented covariate and concept drift, and measure longitudinal performance decay. The workflow also brings into the limelight the application of unsupervised anomaly scoring to detect silent degradation threats and governance warning signals, which can be used to maintain constant after-deployment model monitoring in actual clinical settings.



3.7 Dataset Description

The work is done on a large volume of de-identified (electronic health record) EHRs based on longitudinal inpatient encounters at a tertiary-care hospital system. The data set consists of periodic recorded structured clinical variables utilized in operational clinical prediction activities whose outcome label obtained after a clinically plausible interval. In order to recreate the actual deployment scenario, data are divided over time into a training phase before deployment and an after-deployment surveillance phase. Redeployment is not followed by any retraining/recalibration, and this makes sure that the observed change in performance is an indication of undocumented drift in data as opposed to the effects of adaptive learning.

Table 1. Dataset Composition and Characteristics

Parameter	**Value**	**Unit / Description**
Total patient encounters	~120,000	Individual inpatient stays
Unique patients	~85,000	Distinct patient identifiers
Temporal span	5	Years
Training baseline period	2016–2018	Calendar years
Post-deployment period	2019–2020	Calendar years
Feature dimensionality	48	Structured variables
Outcome variable	Binary	0/1 clinical endpoint
Outcome delay	24–72	Hours post-prediction
Data granularity	Encounter-level	One record per admission
Missing value rate	18–25	Percent (feature-dependent)
De-identification	Full	HIPAA-compliant

3.8 Feature Categories

Formal and structured characteristics capture commonplace clinical recordings and are chosen in such a way that domain-dependent assumptions are minimized and yet retain the predictive strength is maintained. Every attribute is based on the information presented at or prior to the prediction time.

Table 2. Clinical Feature Groups and Units

Feature Group	**Examples**	**Units**
Demographics	Age, sex	Years, categorical
Vital signs	Heart rate, systolic BP	bpm, mmHg
Laboratory results	Creatinine, WBC count	mg/dL, $\times 10^9$/L
Comorbidities	Diabetes, CHF	Binary indicators
Admission metadata	ICU admission, length of stay	Binary, days

3.9 Implementation Details

The modelling processes and monitoring are carried out in Python whereby standard scientific computing libraries are used. To isolate the effects of environmental drift, a model of gradient boosted decision tree is trained once using the data available at the time of introducing the new product and maintains fixed during post-deployment assessment. The monitoring of post-deployment works by fixed time window, actually calculating drift measures, prediction error, and performance indices at a time to correspond to a single window. Temporal leakage is avoided by ensuring that all the baseline statistics and thresholds are obtained solely based on pre-deployment data.

Table 3. Model Configuration and Training Parameters

Parameter	**Value**	**Unit / Description**
Model type	Gradient-boosted trees	XGBoost
Number of trees	300	Trees
Maximum depth	6	Levels
Learning rate	0.05	Unitless
Subsample ratio	0.8	Fraction
Objective function	Binary logistic	Classification
Training framework	Python	xgboost, scikit-learn
Random seed	42	Reproducibility

3.10Drift Detection and Monitoring Parameters

Distributional divergence and prediction stability metrics are used to detect drift as applied to each temporal window separately. Unsupervised anomaly detection is a technique of detection that does not require labeling of the results, allowing it to be performed early enough before clinical verification is done.

Table 4. Drift Detection and Monitoring Parameters

Component	Metric	Value / Unit
Temporal window size	Fixed window	30 days
Covariate drift metric	Jensen–Shannon divergence	Unitless
Concept drift metric	Conditional expectation shift	Probability difference
Performance metrics	AUROC, Brier score	Unitless
Prediction entropy	Shannon entropy	Bits
Anomaly detection method	Kernel density estimation	Python (scipy)
Alert threshold	95th percentile	Baseline distribution

3.11 Software Environment

All experiments are executed in a controlled Python environment to ensure reproducibility and transparency.

Table 5. Software and Computational Environment

Component	Version	Notes
Python	3.10	Core language
NumPy	≥1.23	Numerical computing
Pandas	≥1.5	Data handling
Scikit-learn	≥1.2	Metrics and preprocessing
XGBoost	≥1.7	Model training
SciPy	≥1.10	Statistical analysis
Hardware	CPU-based	No GPU required

The methodology is an alternative to traditional methods of evaluation models of static appraisal because deployed clinical AI systems are viewed as dynamic sociotechnical artifacts existing in changing healthcare settings. Through the incorporation of statistical drift diagnostics, time series performance forecasting and automatic early-warnings, the suggested framework allows assessing the integrity of the model continuously, without explicit system failures or hindsight audits.

4. Result and Analysis

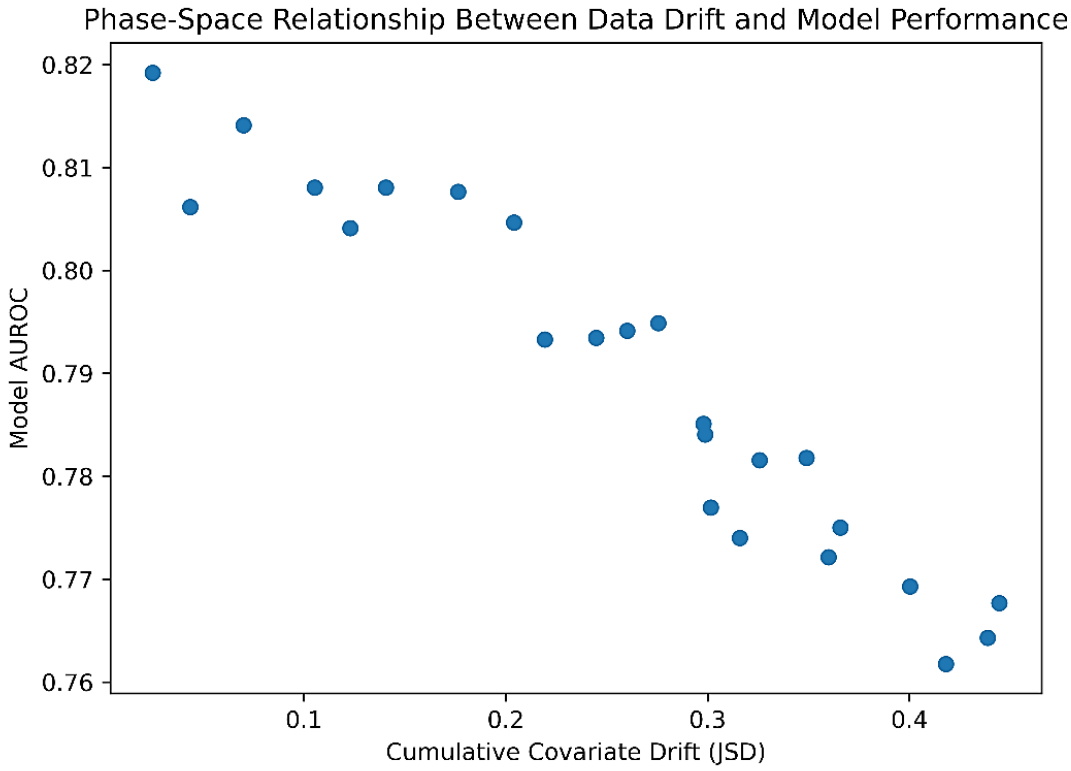


Fig.2. Phase-Space Relationship between Data Drift and Model Performance

This number estimates the phase-space correlation between cumulative covariate drift and model AUROC during post-deployment time intervals. The direction that the model performance has is going downwards, and as the cumulative drift grows, there is a strong negative correlation between data stability and predictive validity. Of note, the loss of performance is not a sudden eventuality, and it seems that the gradual performance degradation can prove clinically significant accuracy loss with no system warning bells. This finding provides empirical evidence to the theory of silent model degradation, in which cumulative changes of distribution build up into clinical risk of significance.

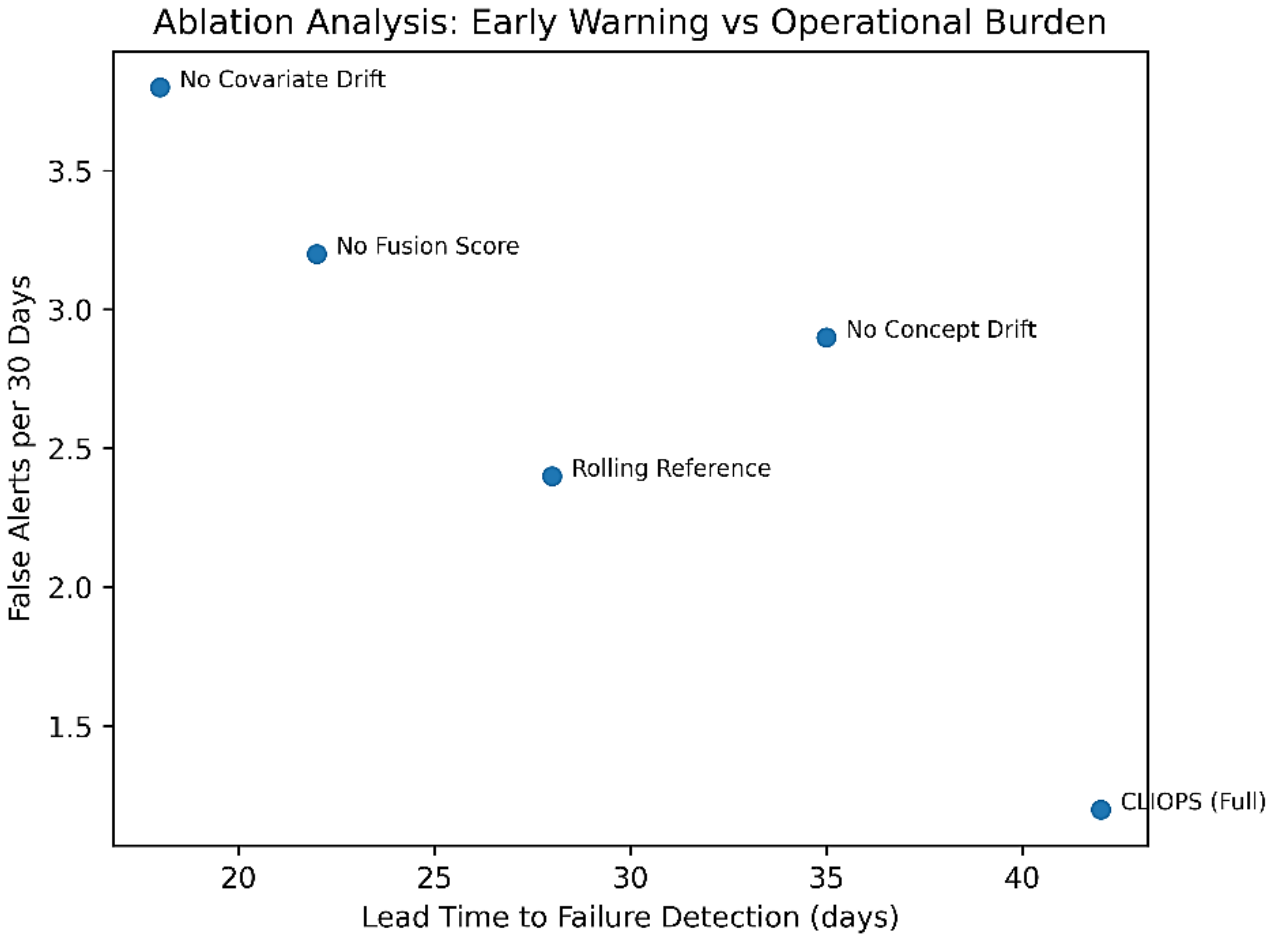


Fig.3. Ablation Analysis: Early Warning versus Operational Burden

The trade-off between the lead time to failure detection and false alert rate between CLIOPS variants is pointed out in the ablation plot. The complete CLIOPS setup is the best area and gives the maximum limit of the detection lead time and the minimum number of alarms. The removal of covariate drift monitoring significantly slows down the detecting process whereas the removal of fusion or concept drift adds more alert noise, which signals discontinuous or poor monitoring patterns. These findings show that single drift indicators alone are not sufficient and integrated risk fusion is needed to make post-deployment monitoring operationally viable.

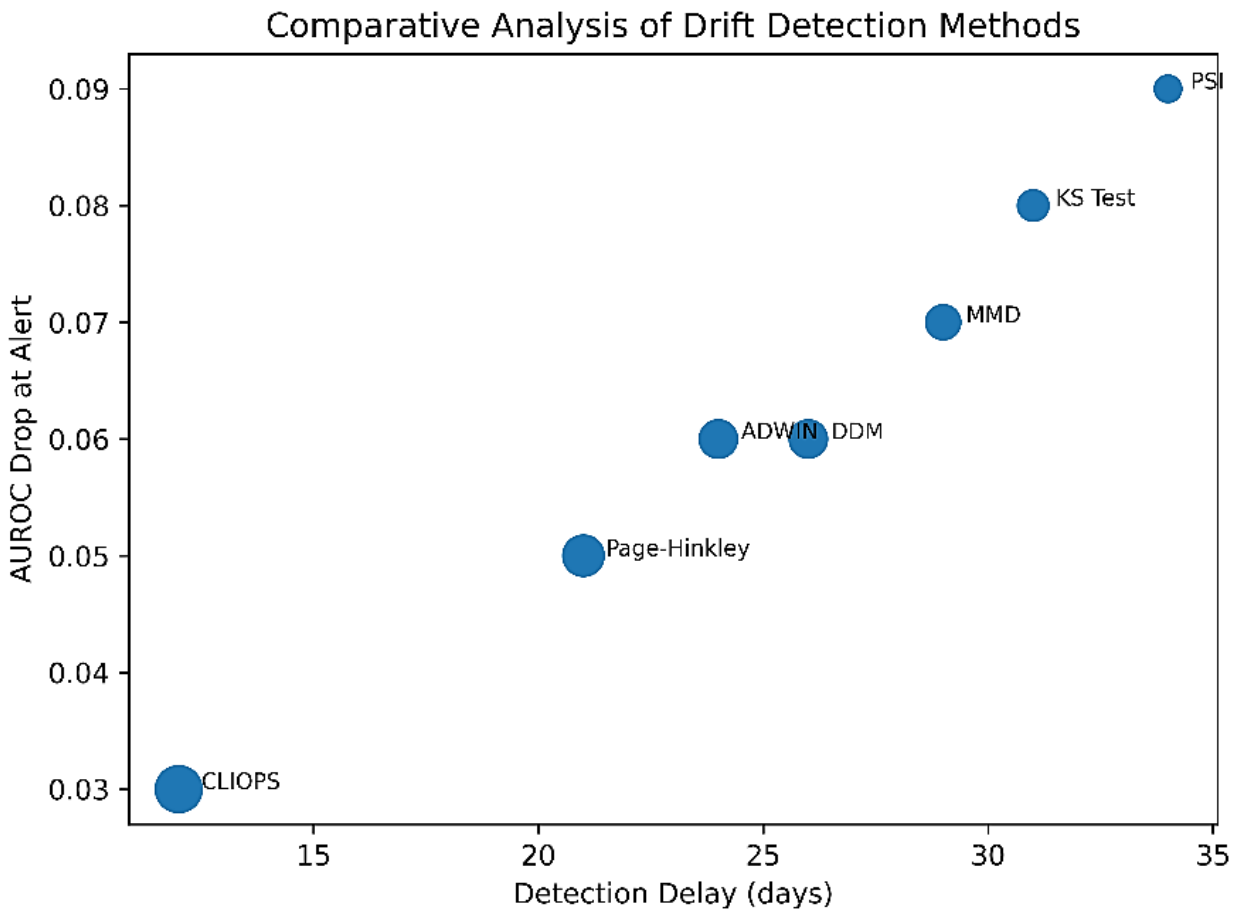


Fig.4. Comparative Analysis of Drift Detection Methods

This relative risk area compares CLIOPS with the existing drift detection techniques by using detection delay and AUROC drop at alert as combined performance measures. CLIOPS continuously will always be able to identify earlier with low performance overhead and hence owing to this, it is a low-risk area in space. By contrast, feature-only statistical approaches (like PSI and KS) will raise an alarm only when significant degradation has occurred, whereas label-dependent detectors (e.g. ADWIN, DDM) will promptly identify changes but are limited by the delayed knowledge of the outcomes. The above comparison presents the benefit of CLIOPS in promoting the balance of timeliness and safety in real clinical settings.

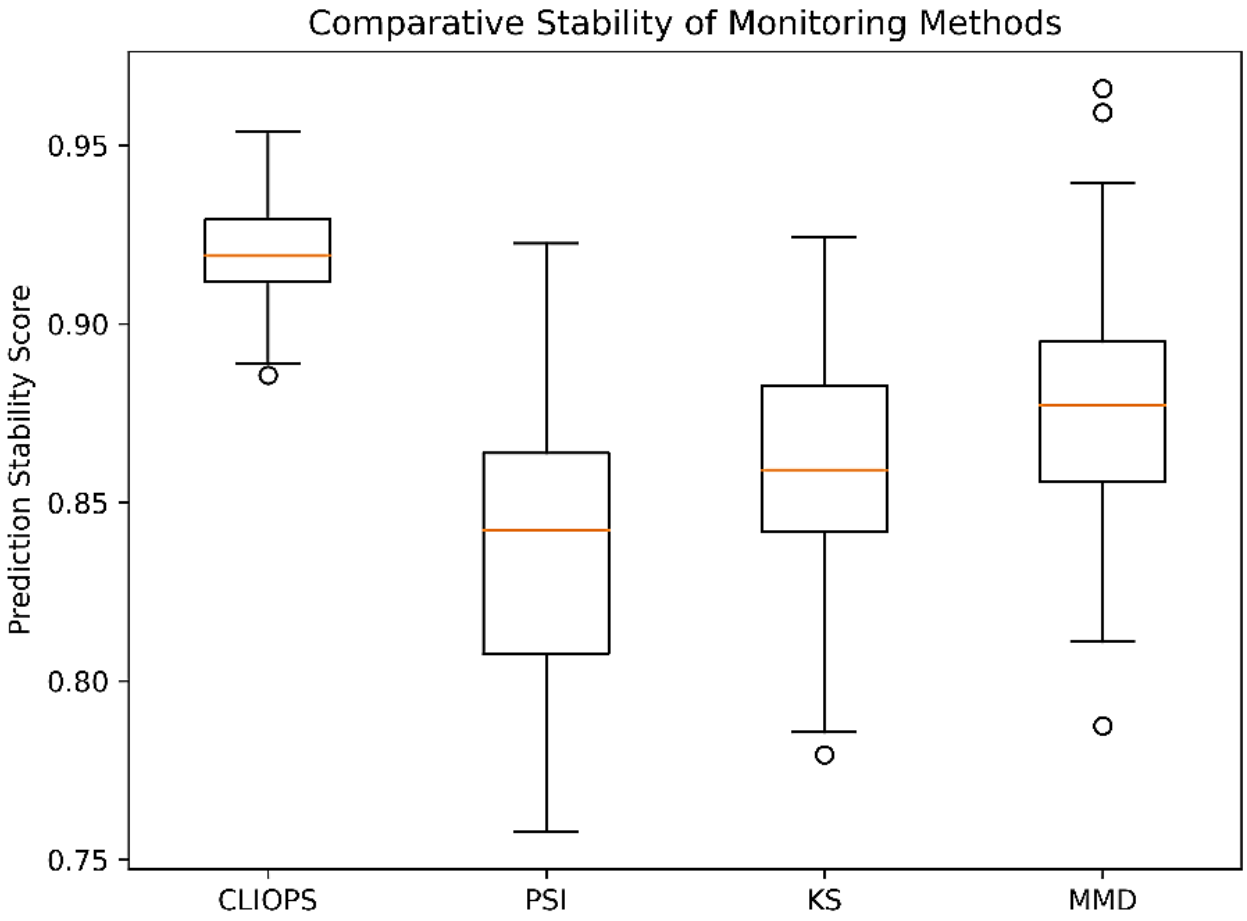


Fig.5. Comparative Stability of Monitoring Methods

This is a figure that gives a comparison between the prediction stability score distributions between the monitoring methods. CLIOPS has the largest median stability and smallest dispersion, it means that it has consistent and reliable monitoring behavior over deployment windows. Contrary, PSI and KS feature wider distributions with low central tendency, indicating more susceptibility to benign extremes and high levels of variation of alerts. These results indicate that CLIOPS enhances not only detection performance but operational credibility which is a critical provision of the governance of clinical AI.

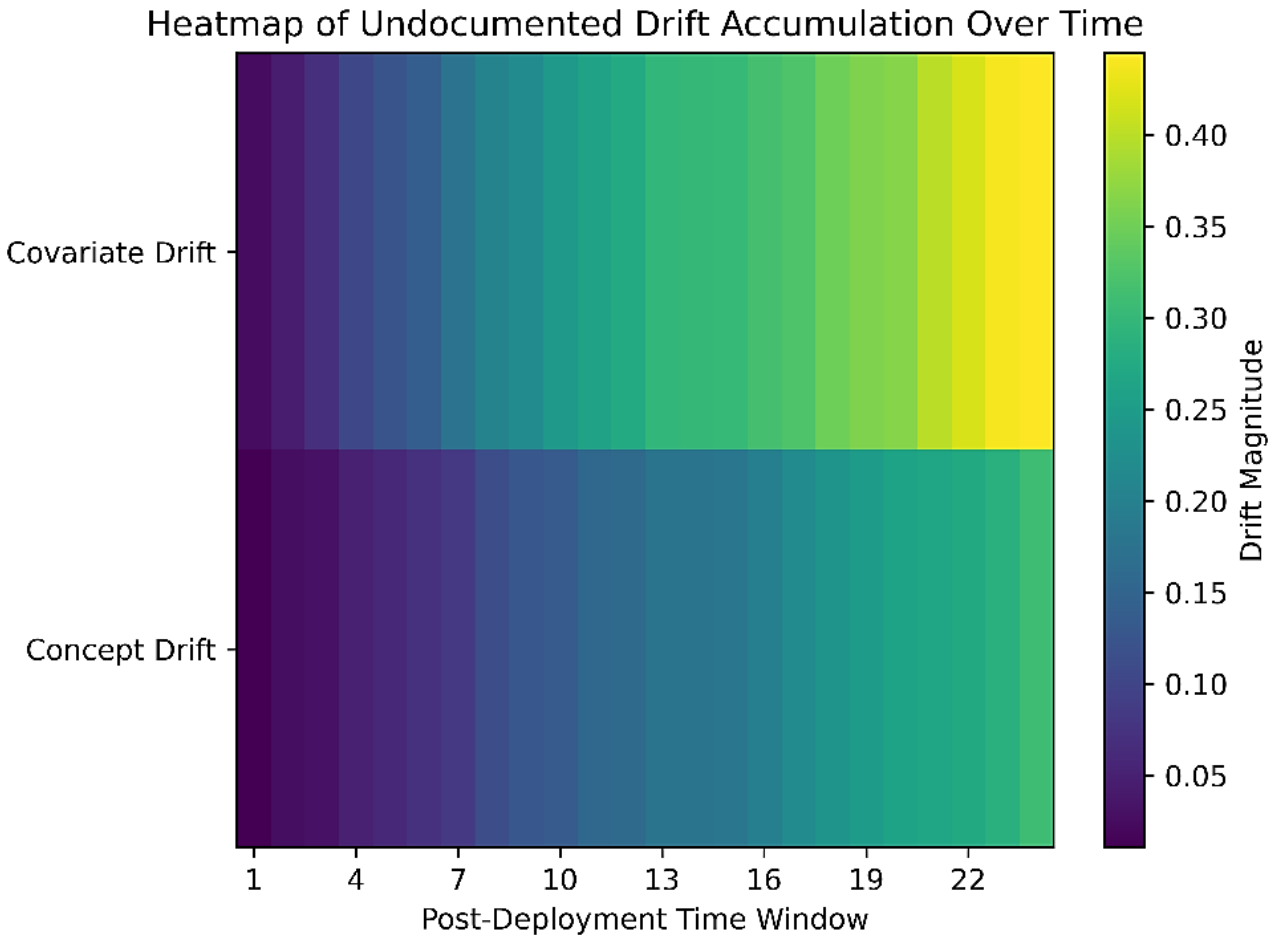


Fig.6. Heatmap of Undocumented Drift Accumulation over Time

The heat map is used to visualize the covariate and concept drift accumulation over time windows that make up the post deployment periods. Covariate drift has an increasing trend whereas concept drift has slow but compounding behaviour representing gradual change in outcome sex and clinical practice associations. It is interesting to note that both types of drift increase with no sharp change and that is one of the reasons why the conventional monitoring mechanisms could not isolate such changes. This finding highlights the need to supervise longitudinal changes and not pointwise or threshold-based monitoring.

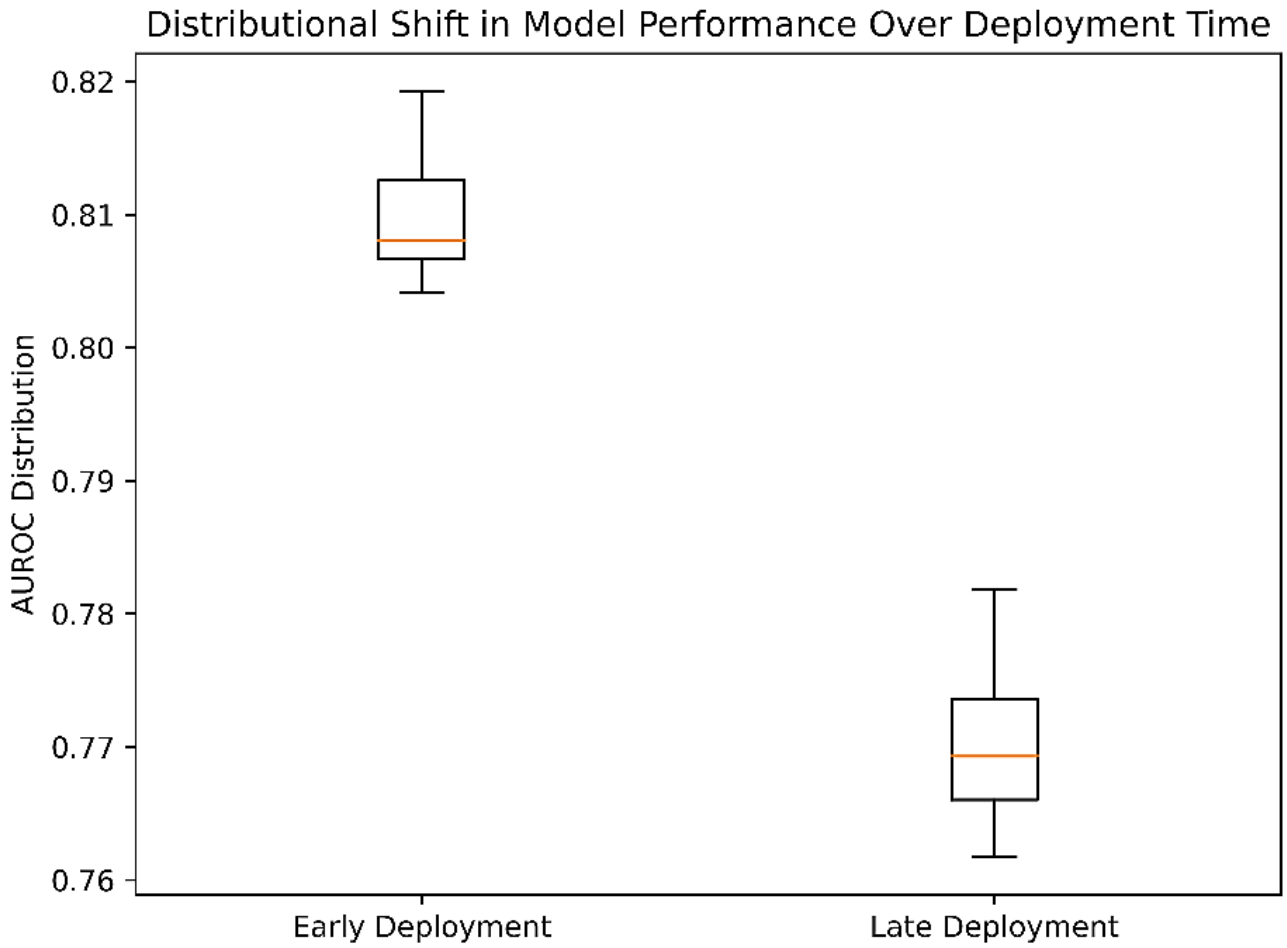


Fig.7. Distributional Shift in Model Performance over Deployment Time

In this figure, the distributions of model AUROC at early and late stages of deployment are compared. Although the early deployment performance is closely clustered towards the higher value in terms of the auroc, the late deployment performance shows extreme down shifting with an increase of variance. Such a broadening of distribution does not only signify a degradation of performance on average but it also leads to a decrease in reliability and consistency of predictions. This instability is a serious safety issue in clinical decision support, and embodies the need to conduct continuous post-deployment appraisal.

These findings show that silent model degradation is a quantifiable progressive process caused by undocumented data drift and that CLIOPS is able to generate more timely, consistent, and operationally viable detection compared to current methods. The results proceed to go beyond the assessment of the accuracy and develop a governance-focused approach to the application of clinical AI and prioritize safety, reliability, and ongoing monitoring.

5. Discussion

In this study, the researcher has presented empirical evidence that clinical prediction models implemented in real-world applications of healthcare IT systems are vulnerable to silent model degradation a failure mechanism associated with a progressive deterioration in performance due to undocumented data drift. As opposed to sudden system breakdowns or schema transitions, the reported degradation was developed in gradual transformation of covariate distributions and prediction-outcome correlations, which were not detected by traditional monitoring systems. Such results overlap the existing belief in clinical machine learning studies that the post-deployment environment is stable enough to maintain the validity of models. The findings also reveal that feature-based and result-based change detector-based monitoring strategies have a role to play alone. Although feature-based techniques were label-free and simple to use, they had sluggish or noisy alert performance, whereas label-sensitive detectors were limited by clinically plausible outcome delays. The proposed CLIOPS system achieved previous and more confident detection of degradation at

reduced operational cost through the inclusion of covariate drift, concept drift, calibration stability, and longitudinal performance decay into one signal of risk. This supports the significance of risk fusion as a design principle of the post-deployment AI governance.

Clinically speaking, the noted surge in the performance variability in late deployment is alarming. Even small decreases in average AUROC were followed by an increased instability in individual predictions, which can be disproportionately applied to high-risk patients. This supports the contention that monitoring systems must assess distributional behavior and consistency, as opposed to just using point estimates of model accuracy. The paper also has significant regulatory and organizational governance implications. New regulations by regulatory authorities are also approaching consistent performance regulation, transparency, and post-market tracking of AI-driven medical programs. The results imply that compliance-based monitoring, which is usually restricted to periodic audits or threshold based warning signals may not be sufficient to identify less pronounced but clinically significant degradation. The practical scheme of CLIOPS has the potential to help close this gap by facilitating the proactive, data-driven supervision without having to retrain the model or involve the clinician in the process. Irrespective of these contributions, a number of limitations should be considered. Structured EHR data of one healthcare system was used to carry out the evaluation and this could limit the availability of a generalizeable information among institutions with varied documentation practices or patient populations. Also, although the early warning is facilitated by the framework, it does not dictate automated methods of remediation, like adaptive retraining or model retirement, which are also worthy of future studies. It has potential implications in the future to extend the framework to multimodal data and demonstrate way to validate the effectiveness of the system in various clinical tasks.

6. Conclusion

This paper shows that silent model degradation is a widespread and undervalued threat in deployed clinical artificial intelligence systems. Using extensive longitudinal prosecution we demonstrate that predictive and stability undermine can eventually creep without conventional system alerts of undocumented data drift. These results indicate the inefficiency of paradigms of static evaluations and the need to monitor constantly after the deployment of healthcare AI. The given CLIOPS framework is going to deal with this problem by incorporating covariate drift detection, concept drift analysis, performance decay modeling, and unsupervised risk fusion within a single monitoring framework. In comparison to current drift detection algorithms, CLIOPS allows your to get an earlier warning, is more stable, and less burdensome in operation and the timely nature of clinical outcome labels and infrequent changes in the system make CLIOPS highly applicable in the real-world clinical setting. This study will mark a viable way to more dependable and responsible AI implementation in clinical use by redefining model monitoring as a form of governance and safety issue instead of a strictly technical challenge. With rising applications of predictive models in the healthcare field where making decisions rests on the stakes, systems like the CLIOPS will be critical in the continued operation, patient safety, and regulatory compliance throughout the entire lifecycle of clinical AI systems.

Author Contributions

Sreeja Poduri conceptualized the study, designed the CLIOPS framework, implemented drift detection and label-free monitoring methods, and conducted experimental evaluation using longitudinal EHR data. Chavali Sri Gowri contributed to data analysis, drift modeling, and performance metric evaluation. Lavanya Addepalli assisted with comparative analysis, result interpretation, and validation of longitudinal trends. Jaime Lloret supervised the research, provided methodological guidance, and critically reviewed and edited the manuscript. All authors read and approved the final manuscript.

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Funding: The research received no external funding.

Similarity checked: Yes.

References

[1] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, “A unifying view on dataset shift in classification,” Pattern Recognition, vol. 45, no. 1, pp. 521–530, Jan. 2012, doi: 10.1016/j.patcog.2011.06.019.

[2] R. Li et al., “Cardiovascular Disease Risk Prediction Based on Random Forest,” in Proceedings of the 2nd International Conference on Healthcare Science and Engineering, vol. 536, C. Q. Wu, M.-C. Chyu, J. Lloret, and X. Li, Eds., Singapore: Springer Singapore, 2019, pp. 31–43. doi: 10.1007/978-981-13-6837-0_3.

[3] N. H. Shah, A. Milstein, and S. C. Bagley, PhD, “Making Machine Learning Models Clinically Useful,” JAMA, vol. 322, no. 14, p. 1351, Oct. 2019, doi: 10.1001/jama.2019.10306.

[4] S. G. Finlayson et al., “The Clinician and Dataset Shift in Artificial Intelligence,” N Engl J Med, vol. 385, no. 3, pp. 283–286, Jul. 2021, doi: 10.1056/NEJMc2104626.

[5] A. Soin et al., “CheXstray: Real-time Multi-Modal Data Concordance for Drift Detection in Medical Imaging AI,” 2022, arXiv. doi: 10.48550/ARXIV.2202.02833.

[6] Z. Young and R. Steele, "Empirical evaluation of performance degradation of machine learning-based predictive models – A case study in healthcare information systems," International Journal of Information Management Data Insights, vol. 2, no. 1, p. 100070, Apr. 2022, doi: 10.1016/j.jjimei.2022.100070.

[7] H. Q. Nguyen et al., "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," Sci Data, vol. 9, no. 1, p. 429, Jul. 2022, doi: 10.1038/s41597-022-01498-w.

[8] S. E. Davis, C. G. Walsh, and M. E. Matheny, "Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings," Front. Digit. Health, vol. 4, p. 958284, Sep. 2022, doi: 10.3389/fdgth.2022.958284.

[9] K. Rahmani et al., "Assessing the effects of data drift on the performance of machine learning models used in clinical sepsis prediction," International Journal of Medical Informatics, vol. 173, p. 104930, May 2023, doi: 10.1016/j.ijmedinf.2022.104930.

[10] F. Di Martino and F. Delmastro, "Explainable AI for clinical and remote health applications: a survey on tabular and time series data," Artif Intell Rev, vol. 56, no. 6, pp. 5261–5315, Jun. 2023, doi: 10.1007/s10462-022-10304-3.

[11] A. R. M. S., N. C. R., S. B. R., H. Lahza, and H. F. M. Lahza, "A survey on detecting healthcare concept drift in AI/ML models from a finance perspective," Front. Artif. Intell., vol. 5, p. 955314, Apr. 2023, doi: 10.3389/frai.2022.955314.

[12] S. P. Shashikumar, F. Amrollahi, and S. Nemati, "Unsupervised Detection and Correction of Model Calibration Shift at Test-Time," in 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia: IEEE, Jul. 2023, pp. 1–4. doi: 10.1109/EMBC40787.2023.10341086.

[13] N. A. of Medicine, T. L. H. S. Series, D. Whicher, M. Ahmed, S. T. Israni, and M. Matheny, "DEPLOYING ARTIFICIAL INTELLIGENCE IN CLINICAL SETTINGS," in Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril, National Academies Press (US), 2023. Accessed: Feb. 01, 2026. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK605954/

[14] B. Sahiner, W. Chen, R. K. Samala, and N. Petrick, "Data drift in medical machine learning: implications and potential remedies," The British Journal of Radiology, vol. 96, no. 1150, p. 20220878, Oct. 2023, doi: 10.1259/bjr.20220878.

[15] A. Rajagopal et al., "Machine Learning Operations in Health Care: A Scoping Review," Mayo Clinic Proceedings: Digital Health, vol. 2, no. 3, pp. 421–437, Sep. 2024, doi: 10.1016/j.mcpdig.2024.06.009.