A Method for Automating Iterative Machine Learning Model Retraining on Event Streams in Decision Support Systems

Arseniy Lomakin[1,2(✉)] and Aleksandr Makarov[1]

[1] Volgograd State Technical University, Lenin avenue, 28, Volgograd 400005, Russia
arseny.lomakin@gmail.com, a.o.makarov2003@gmail.com
[2] Volgograd State Medical University, Fallen Fighters Square, 1, Volgograd 400066, Russia

Abstract. Clinical Decision Support Systems (CDSS) increasingly rely on machine learning models that suffer performance degradation under concept drift in streaming clinical data. Existing approaches typically address drift detection, anomaly identification, or continual learning in isolation, lacking a unified framework with formal non-degradation safeguard. This paper proposes an event-driven continual learning method (Method ECLA) for automated lifecycle management of classifier ensembles in CDSS. The approach integrates Mahalanobis distance-based anomaly detection, clustering-driven micro-batch accumulation, and stability-preserving local adaptation via Elastic Weight Consolidation and Experience Replay. A formal non-degradation constraint, enforced via an automatic rollback mechanism, ensures that model updates do not reduce predictive performance beyond a predefined tolerance ($\varepsilon = 0.05$). The method is evaluated on 70,421 real ophthalmological cases over a 180-day operational horizon across 10 independent runs. Experimental results demonstrate that Method ECLA (Event-driven Continual Learning with Anomaly clustering) outperforms static and periodically retrained baselines, achieving Accuracy of 0.890 and AUC-ROC of 0.873, while reducing Time-to-Recovery (TTR) to 1.2 days and maintaining computational efficiency independent of dataset size. The proposed Method ECLA enables stable, adaptive, and resource-efficient deployment of machine learning models in real-world CDSS environments.

Keywords: Clinical Decision Support Systems · Continual Learning · Concept Drift · Anomaly Detection · Ensemble Learning · IOL Power Calculation

1 Introduction

Machine learning models are increasingly being integrated into Clinical Decision Support Systems (CDSS) to help physicians make more accurate diagnoses and select the appropriate treatment plan. Such systems use large volumes of clinical data to train models that assist in medical decision-making.

However, machine learning models implemented in Clinical Decision Support Systems (CDSS) in real-world medical practice are inevitably subject to obsolescence.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16769, pp. 292–308, 2027.
https://doi.org/10.1007/978-3-032-30494-0_19

Advances in diagnostic methods in modern medicine lead to the identification of previously unknown diseases and the refinement of classifications of existing nosological forms, which allows for the development of more precise treatment plans for each individual patient. This gives rise to the phenomenon of conceptual drift [1–3], in which the statistical relationship between input features (e.g., patient condition data) and target variables (e.g., the proposed diagnosis) changes, which over time leads to a decrease in the accuracy and predictive value of machine learning models. As a result, the likelihood of incorrect system recommendations increases, which potentially raises the risk of medical errors and, consequently, jeopardizes the quality of patient care.

Traditional approaches to maintaining the quality of machine learning models are based on periodic retraining. Scheduled retraining involves updating models on new datasets according to a predefined schedule, including instances of full retraining using the entire available dataset. Despite its relative simplicity of implementation, this practice cannot ensure effective adaptation to the actual flow of clinical data. During periods of stability in the data distribution, computational resources are wasted on scheduled retraining cycles. However, in situations of sudden shifts in the distribution of input data, the model remains outdated until the scheduled retraining time arrives, and in the case of full retraining, excessive computational costs and a decrease in accuracy on outliers are observed.

Traditional approaches include hyperparameter optimization using GridSearch. This method allows determining the optimal model configuration on a fixed dataset; however, it does not provide mechanisms for responding to outliers and does not include a trigger threshold for initiating retraining upon detection of conceptual drift, thereby requiring manual retriggering.

Another traditional approach is the incremental neural network (INN) approach with step-by-step training. Despite its ability to gradually learn patterns, it exhibits low accuracy and requires manual retraining whenever new data is accumulated.

One alternative approach is the online updating of model weights, which ensures adaptation to the changing distribution of clinical data as each new example is received. However, such retraining carries the risk of catastrophic forgetting—a phenomenon in which the model loses previously learned patterns when trained on new data [4].

The authors set out to automate the iterative training of machine learning models in Clinical Decision Support Systems (CDSS) based on an event stream by applying the developed Method ECLA, which reduces the resource costs of retraining as the data volume increases while adhering to three key criteria: a safeguard of no degradation; reactivity—response upon the occurrence of an anomaly; and adaptability—dynamic expansion of recognizable patterns.

This article discusses the stages of creating and implementing the proposed Method ECLA, evaluates its key properties, conducts a comparative analysis with existing approaches, and presents a software tool developed for this method.

The contributions of this work are as follows:

- An event-driven continual learning framework for CDSS that unifies anomaly detection, micro-batch accumulation, and stability-preserving local adaptation under a formal non-degradation safeguard;

- A statistically principled anomaly clustering mechanism that forms representative micro-batches without dependence on the total dataset size, ensuring constant-cost retraining;
- A dual-memory adaptation strategy combining Elastic Weight Consolidation and Experience Replay, which prevents catastrophic forgetting while learning new clinical patterns;
- Comprehensive experimental validation on 70,421 ophthalmological cases, demonstrating stable performance over 180 days and significant improvements over currently deployed retraining strategies.

2 Methodology

2.1 Proposed Solution

In existing research on the application of machine learning in CDSS, adaptation to conceptual drift, anomaly detection, and continuous learning are typically addressed in isolation [5–7]. However, in real-world operating conditions, CDSS function in non-stationary environments where these problems may arise simultaneously [8, 9]. Consequently, effective long-term implementation requires a unified approach that ensures: 1) timely detection of shifts in data distribution; 2) stable and controlled model updates; 3) resistance to catastrophic forgetting of previously learned patterns; 4) a safeguard that performance remains within a specified threshold during model adaptation.

Based on the literature review, it was found that existing approaches often fulfill only a few of these requirements [10–12]. The absence of a unified formalized structure that ensures the simultaneous fulfillment of all four requirements justified the need to develop and experimentally validate the proposed method ECLA.

In practice, clinical data in CDSS arrive sequentially and can be formalized as a stream of paired observations (1):

$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i)\} \tag{1}$$

where $x_i \in \mathbb{R}^d$ denotes the feature vector of the i-th clinical case, and $y_i \in \mathbb{R}$ denotes the corresponding target variable. This formulation reflects the continuous accumulation of clinical knowledge and requires incremental, controlled model updates under conditions of potential distribution shifts.

Let the ensemble of machine learning models deployed in the CDSS at time t be defined as (2):

$$M_t = \left\{M_t^1, M_t^2, \ldots, M_t^n\right\} \tag{2}$$

where each model $M_t^i (i = 1, 2, \ldots, n)$ performs a classification task aimed at selecting the most appropriate formula (Barrett, Haigis, Hoffer Q, Holladay) for calculating the power of an intraocular lens (IOL) for a given clinical case.

The ensemble's performance is evaluated using the quality function $Q(M_t)$, which reflects the predictive ability of the entire ensemble of models on a fixed dataset at time t. In practical applications, this function can be interpreted as standard classification quality metrics or AUC-ROC.

From the perspective of the machine learning lifecycle [13], CDSSs are dynamic systems that require constant maintenance of the relevance of predictive models. The main task lies not in initial training, but in controlled adaptation to new clinical data over time. In this regard, an appropriate mechanism for updating machine learning models is required, one that is not prone to degradation and instability during periods of changing data distribution.

Method ECLA formalizes a constraint on the preservation of the model's predictive performance, defined as (3)

$$Q(M_{t+1}) \geq Q(M_t) - \varepsilon, \tag{3}$$

where $Q(M_t)$ is a scalar measure of model quality at iteration t, calculated based on standard classification metrics (Accuracy, AUC-ROC, F1-score), and ε ($\varepsilon = 0.05$) is a predefined tolerance parameter, empirically selected to balance model adaptability with performance stability during data updates. In this study, Accuracy and AUC-ROC are used as the primary metrics for evaluating predictive capabilities, which is consistent with common practice in evaluating machine learning models in medical applications [14].

Thus, the goal of the proposed method is to automatically manage the lifecycle of machine learning models in a clinical data stream while simultaneously meeting key requirements: responsiveness, adaptability, and quality preservation. To achieve this goal, an event-driven method for iterative ensemble learning of classifiers has been developed to support clinical decision-making when selecting the optimal formula for calculating the optical power of an IOL replacing a clouded lens.

It should be noted that the proposed method is not limited solely to cataract surgery. The basic framework—event-driven clustering of anomalies and local fine-tuning with a non-degradation safeguard—is applicable to any CDSS in which an ensemble of classifiers operates on a data stream subject to conceptual drift. The scope of application can be extended to other surgical specialties, diagnostic tasks, and medical information systems.

2.2 Anomaly Detection Mechanism

One of the key components of Method ECLA is the identification of clinical observations that deviate from the statistical distribution of the training dataset. The occurrence of such observations may indicate the emergence of new patterns, so the CDSS must respond adaptively to such cases.

To detect novelty in the incoming stream of clinical data, Method ECLA uses a metric such as the Mahalanobis distance, which measures the statistical deviation of a new observation from the multidimensional distribution of the training dataset [15].

A new clinical case, represented by the feature vector x, is considered an anomalous event e_a if condition (4) is satisfied:

$$D_m(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)} > \tau \tag{4}$$

where μ is the sample mean of the feature vector of the training sample, Σ is the covariance matrix of the feature space, and τ is the novelty detection threshold.

The choice of the Mahalanobis distance is based on the statistical properties of the ophthalmic parameters used in IOL calculations: gender (GENDER) and the eye being examined (EYE); biometric parameters of the eye—axial length (AL), anterior chamber depth (ACD), lens thickness (LT), white-to-white distance (WTW); keratometric parameters of the anterior surface (K1, K2, ΔK), total (TK1, TK2, ΔTK), and posterior surface (PK1, PK2, ΔPK2); corneal thickness (CCT), Kappa angle (CW-Chord), and refractometry (AxR); calculated values of the spherical equivalent and optical power of the IOL using four formulas: Barrett, Haigis, Hoffer Q, Holladay. The listed characteristics are mutually correlated and measured on different scales and in different units of measurement.

Under such conditions, the Euclidean distance becomes statistically incorrect, as it assumes the independence and equal-scalability of the features. However, the Mahalanobis distance explicitly accounts for the covariance structure via the matrix Σ^{-1}, which makes it invariant to the scales and correlations of the features.

The novelty detection threshold τ is derived mathematically. Assuming that the training data follow a multivariate normal distribution—which is approximated by an empirical estimate of the covariance matrix—the square of the Mahalanobis distance follows a chi-squared distribution with k degrees of freedom (5):

$$D_m^2(x) \sim \chi_k^2 \tag{5}$$

where k denotes the dimension of the feature space.

Consequently, the threshold value τ is chosen as (6):

$$\tau = \sqrt{\chi^2_{k,1-a}} \tag{6}$$

where α denotes the significance level, which determines the probability of a false positive detection of an anomaly in the input data.

In this study, the feature space has a dimension of $k = 33$, corresponding to 32 biometric parameters measured by ophthalmic devices (IOL Master 700, Tomey OA-2000, Canon RK-1, Sirius, Pentacam). At a significance level of $\alpha = 0.01$, the obtained threshold is $\tau \approx 4.82$. This means that no more than 1% of standard clinical observations will be incorrectly classified as anomalies, thus maintaining the false positive rate below 1%.

It is worth noting that the detected anomalous data are not used immediately to retrain the model. Instead, they are accumulated and subjected to further processing using the clustering mechanism described in the following subsection.

2.3 Anomaly Clustering and Micro-batch Formation

The clustering scheme for newly detected patterns, implemented in Method ECLA, helps avoid instability caused by rare isolated anomalies that may occur in incoming streams of medical data. In other words, this approach allows the system to accumulate a statistically significant number of anomalous observations before beginning model adaptation.

Each anomalous observation x_i is routed to the nearest cluster of anomalies according to the Mahalanobis distance. The optimal cluster is defined as (7):

$$C_{opt} = \text{argmin}_k D_m(x_i, \text{centroid}(C_k)) \tag{7}$$

where $\text{centroid}(C_k)$ denotes the centroid of the kth cluster of anomalies.

To classify incoming medical data, the Mahalanobis distance from the current observation x_i to the center of the nearest cluster C_{opt} is calculated. If the distance between x_i and C_{opt} satisfies condition (8):

$$D_m(x_i, \text{centroid}(C_{opt})) \leq \gamma \tag{8}$$

where γ denotes the decision threshold that distinguishes the region of acceptable statistical deviations in the data from the region of novelty; in this case, the method interprets the observation as belonging to a previously detected anomaly type, and it is added to the corresponding cluster. Otherwise, this indicates the emergence of a previously undetected clinical pattern, and a new cluster of anomalies is created (9).

$$C = C \cup \{\{x_i\}\} \tag{9}$$

Model retraining does not start immediately after a new cluster is formed. Instead, the system accumulates outliers until the cluster itself reaches a predefined minimum size.

Method ECLA includes the local retraining procedure LearnLocalExpert(C_{opt}), which is triggered if condition (10) is satisfied:

$$|C_{opt}| \geq N_{th} \tag{10}$$

where N_{th} is the threshold value for the minimum cluster size.

The lower bound of the micro-batch threshold $N_{th} = 33$ is justified by the condition of non-degeneracy of the covariance matrix. In a d-dimensional feature space, at least $d + 1$ observations are required for a stable estimation of the covariance matrix. With 32 parameters measured by ophthalmic devices, $d = 32$, from which it follows that the minimum required number of observations is 33. In turn, the upper bound of the micro-batch threshold $N_{th} = 50$ was determined based on an analysis of continuous learning metrics. The theoretical dependence of the Forgetting Measure (FM) and Forward Transfer (FWT) [16] on the batch size N is shown in Fig. 1.

Thus, this interval $N_{th} \in [33, 50]$ represents a scientifically validated optimal range that simultaneously minimizes the FM index and maximizes the FWT index, thereby ensuring a balance between the stability of previously learned patterns and the model's ability to adapt.

2.4 Continual Learning and Catastrophic Forgetting Prevention

The local retraining procedure LearnLocalExpert(C_{opt}), which was introduced and described in the previous subsection, should take into account newly discovered clinical patterns without compromising the model's predictive performance for previously

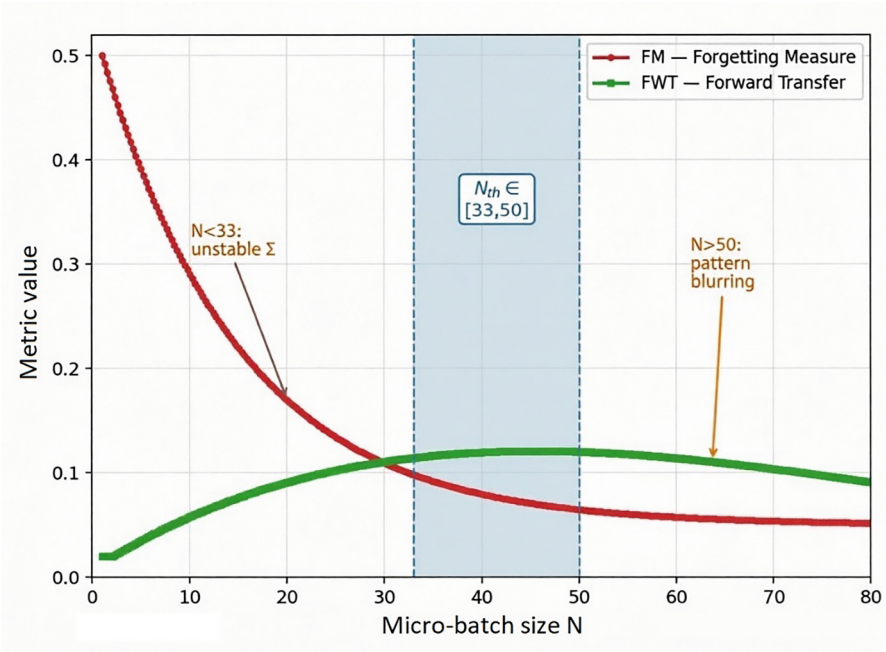

Fig. 1. The dependence of the Forgetting Measure (FM) and Forward Transfer (FWT) on the size of the micro-batch N.

recorded observations. It should be noted that this procedure does not involve creating a new model, but rather retraining the current ensemble of classifiers M_t by fine-tuning its weights. In continuous learning, simple retraining on newly acquired data can lead to catastrophic forgetting, resulting in the loss of previously learned patterns when adapting to new cases. To avoid this, in Method ECLA this problem is solved using two complementary mechanisms, namely Elastic Weight Consolidation (EWC) [17] and Experience Replay [18].

The EWC approach adds a regularization term to the standard loss function, "locking in" the values of weights that are critically important for historical data. The importance of each weight, in turn, is estimated via the diagonal elements of the Fisher matrix, which allows for differentiated penalization of parameter changes. The resulting loss function takes the form (11):

$$L_{EWC}(\theta) = L_B(\theta) + \frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_i^*)^2 \quad (11)$$

where $L_B(\theta)$ denotes the loss function calculated on the current data batch, θ_i^* represents the optimal model weights obtained on the training dataset D, F_i are the diagonal elements of the Fisher matrix, which quantitatively reflect the importance of each weight parameter for preserving predictive power on previously studied distributed data, and λ is the regularization coefficient that controls the balance between learning on new data and preserving historical knowledge. Thus, EWC selectively constrains important weights, allowing plasticity in less critical parameters.

The Fisher information matrix is estimated as the mathematical expectation of the square of the gradient of the log-likelihood with respect to the model parameters (12). This approach quantitatively measures the sensitivity of the likelihood function to changes in each weight. Elements with a high value of F_i correspond to parameters that strongly influence the quality of the input data—changes to them during fine-tuning are penalized more heavily.

$$F_i = E\left[\left(\frac{\partial \log p(D|\theta)}{\partial \theta_i}\right)^2\right] \tag{12}$$

In the proposed system, the EWC mechanism is based on an optimized C++ implementation integrated into the PyTorch framework; since the Fisher matrix is computed based on the log-likelihood gradients for all historical data, its full evaluation requires significant computational resources in a naive implementation.

The second mechanism used in the retraining procedure is Experience Replay, which forms a training batch as a combination of a new anomaly cluster C_{opt} and a random sample from the historical data buffer $\text{Buffer}(D, n_{replay})$ (13):

$$B_{train} = C_{opt} \cup \text{Buffer}(D, n_{replay}) \tag{13}$$

The fact that historical examples are present in every fine-tuning batch ensures that the model continues to analyze standard clinical cases while specializing in the new pattern.

Thus, with naive fine-tuning on only C_{opt}, the accuracy on historical data degrades to Accuracy $\approx$ 0.20. In turn, the method described in this subsection maintains the quality of historical data above the threshold value $\varepsilon = 0.05$, thereby satisfying the non-degradation criterion.

2.5 Event-Driven Iterative Learning Workflow

The mechanisms described above constitute the key components of Method ECLA. This subsection presents a formalized workflow that governs the model's lifecycle in the context of clinical data stream processing. Algorithm 1 shows the proposed method in the formal form.

Algorithm 1 Event-driven iterative learning (Method ECLA)

Input: Initial ensemble M_t, clinical data stream $D = \{(x_i, y_i)\}$, novelty detection threshold τ, cluster acceptance threshold γ, micro-batch size $N_{th} \in [33, 50]$, non-degradation threshold ε
Output: Updated M ensemble

```
Initialize anomaly cluster set C ← ∅
Initialize current model M ← M_t
for each incoming observation x_i ∈ D do
    // --- Event e_a: Anomaly detection ---
    Compute D_m(x_i)
    if D_m(x_i) > τ then
        e ← e_a
        // --- Clustering and accumulation ---
        Find C_opt = argmin_k D_m(x_i, centroid(C_k))
        if D_m(x_i, centroid(C_opt)) ≤ γ then
            C_opt ← C_opt ∪ {x_i}
        else
            Create new cluster C_new = {x_i}
            C ← C ∪ {C_new}
            C_opt ← C_new
        end if
        // --- Event e_th: Micro-batch accumulation ---
        if |C_opt| ≥ N_th then
            M' ← LearnLocalExpert(M, C_opt)
            // --- Non-degradation check ---
            if Q(M') ≥ Q(M) − ε then
                M ← M'
            else
                M ← Rollback(M)
            end if
        end if
    else
        // --- Event e_t: Time-based retraining and Rollback ---
        if TimeIntervalExceeded() then
            M' ← GlobalRetrain(M, D)
            if Q(M') ≥ Q(M) − ε then
```

The operation of Method ECLA is determined by a continuous stream of medical data events (14) that enter the database after processing the biometric parameters measured by ophthalmic devices.

$$E = \{e_1, e_2, \ldots, e_n\} \tag{14}$$

In the context of the proposed solution, three types of events were defined: anomaly event e_a—detection of a clinical case that statistically deviates from the distribution of data learned during the training phase; the data accumulation event e_{th}—the anomaly cluster reaching the minimum micro-batch size required for retraining; the time interval event e_t—the expiration of a specified time interval since the last update.

The developed method formalizes a hierarchy of event priorities: e_a, then e_{th}, then e_t, which is justified by an empirical analysis of continuous learning metrics.

An anomaly event has the highest priority, as it signals the emergence of a previously unobserved clinical pattern requiring an immediate system response. In this case, the newly observed data is processed by the anomaly clustering mechanism, described in detail in Sect. 2.3.

When the size of the anomaly cluster reaches a set minimum threshold, a data accumulation event occurs. This triggers the local model adaptation procedure LearnLocalExpert(C_{opt}), which is described in Sect. 2.4—this is the system's primary adaptation mechanism.

The time interval event serves as a safety mechanism for Method ECLA. If, within a specified time, no anomaly cluster has reached the minimum N_{th} threshold, the system performs global retraining on the entire dataset currently available.

This hierarchy ensures the optimal use of computational resources: computationally intensive global retraining is performed only, when necessary, while local adaptations to new patterns are initiated as data accumulates.

After each retraining procedure, the updated ensemble M'_{t+1} is evaluated using a quality function. The system's transition from state S_t to state S'_t is permitted only if condition (15) is satisfied:

$$Q\left(M'_{t+1}\right) \geq Q(M_t) - \varepsilon \tag{15}$$

This approach ensures that the system's forecasting quality remains within a specified threshold for acceptable degradation.

In addition to the quality degradation limit, Method ECLA implements an automatic Rollback(M_t). If the updated ensemble does not meet the quality criterion, the system automatically restores the previous state of the model, preventing the deployment of versions with degraded predictive performance.

The complete algorithm for Method ECLA—from event registration in the data stream to updating the classifier ensemble or performing a rollback when the non-degradation criterion is violated—is shown in Fig. 2.

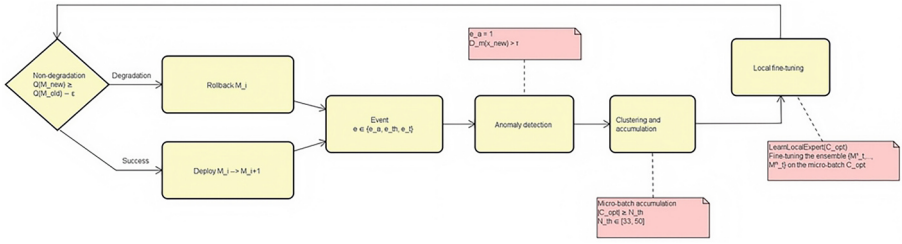

Fig. 2. Flowchart of Method ECLA for automatic iterative learning.

The diagram illustrates a three-level hierarchy for event processing, an anomaly clustering mechanism, quality verification of the updated model, and an automatic rollback procedure.

Thus, the architecture of the developed Method ECLA ensures stable and resource-efficient continuous learning in CDSS under conditions of non-stationary clinical data distributions.

3 Results

3.1 Experimental Setup

The authors conducted an experimental validation of Method ECLA to assess its effectiveness in a CDSS environment. The study utilized an ensemble of classifiers from the "LensCalc" system, designed to assist ophthalmologists in calculating IOLs based on biometric parameters.

The experiment utilized a dataset comprising $|D| = 70\,421$ clinical cases (MNTK, 2018–2025), divided using a stratified method into a training set D_{train} (80%) and a test set D_{test} (20%) subsets. The test set D_{test} remained unchanged across all experiments to ensure the comparability of results.

The extraction of clinical data was automated using an OCR parser (PyTesseract) to process diagnostic reports in PDF format. The raw data, stored in sixth normal form (6NF), was converted to third normal form (3NF) by consolidating one record per patient and aggregating measurements from multiple devices. For duplicate measurements, priority was given to values from devices with higher accuracy; otherwise, the most recent measurement was retained. The input data for the ensemble of classifiers includes 32 clinical parameters described in Sect. 2.2.

Each classifier in the ensemble generates a binary prediction indicating whether a specific IOL calculation formula is applicable (1) or not (0). Ground-truth labels were determined retrospectively: a formula was considered applicable if the postoperative refractive error deviated from the target by no more than ± 0.50 diopters.

To assess the stability of the method, $n = 10$ independent runs were performed with random initial ensemble states (initial accuracy $Acc \in [0.77, 0.83]$) over a horizon of 100 iterations, corresponding to 180 days of system operation.

The initial states for all experiments are shown in Table 1.

Table 1. Initial conditions M_0 for 10 independent experiments.

Parameter	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}
Accuracy M_0	0.79	0.83	0.81	0.77	0.82	0.80	0.78	0.83	0.79	0.81
AUC M_0	0.80	0.84	0.82	0.78	0.83	0.81	0.79	0.84	0.80	0.82

This experimental setup ensures the long-term stability of Method ECLA under conditions of gradual data accumulation and shifts in the distribution.

3.2 Basic Retraining Methods for Comparison

To evaluate the effectiveness of the proposed method, we compared it with three baseline retraining methods widely used in practical machine learning systems.

The first (B_1) is periodic manual retraining at 14-day intervals. This approach reflects the traditional CDSS workflow, where updates are performed infrequently due to high computational costs and the need for manual validation.

The second baseline method (B_2) is automatic scheduled retraining at 7-day intervals without data distribution analysis, using the Cron job scheduler. This approach reduces the need for manual intervention but does not adapt to conceptual data drift.

The third baseline method (B_3) is automatic retraining using GridSearch every 7 days. This approach ensures full adaptation of the model's hyperparameters based on new data, but requires significant computational resources due to the exhaustive search of all combinations.

All methods were evaluated under the identical experimental conditions described in the previous subsection, ensuring the comparability of results, and differences in performance are determined solely by the retraining strategy.

These baselines reflect prevailing maintenance strategies in medical information systems; advanced streaming and continual learning extensions are addressed in the limitations.

3.3 Comparison Results

A comparative analysis of the developed Method ECLA and the baseline retraining methods described in Sect. 3.2 was conducted in accordance with the experimental setup defined earlier in Sect. 3.1.

The evaluation included both measures of predictive performance and measures of the method's stability over time. Prediction quality was assessed using the Accuracy and AUC-ROC metrics, while performance stability during long-term operation was evaluated using the FM and Time-to-Recovery (TTR) metrics.

The Forgetting Measure quantitatively assesses the level of performance degradation accumulated during iterative model updates and is defined as follows (16):

$$FM = \max_{t<T}[Q(M_t) - Q(M_T)] \tag{16}$$

where $Q(M_t)$ represents the value of the quality metric at iteration t, and T corresponds to the final iteration of the experiment.

The TTR metric characterizes the time interval required for the system to recover its predictive capability after a degradation event and is calculated using the following formula (17):

$$TTR = t_{recovery} - t_{event} \tag{17}$$

where t_{event} denotes the specific time at which a performance degradation is observed, and $t_{recovery}$ corresponds to the moment when the model's performance returns to its pre-t_{event} state.

To assess the statistical significance of the identified differences in the predictive ability of the methods, the nonparametric Mann-Whitney U-test [19] was applied for independent samples (sample size $n = 10$). The formulation of the hypotheses to be tested is given by (18):

$$H_0: Q(M_t) = Q(M_0), H_1: Q(M_t) > Q(M_0) \tag{18}$$

at a significance level of $\alpha = 0.05$. The summary results of the experiment are presented in Table 2.

Table 2. Results of a comparative experiment (180-day time horizon, n = 10 independent runs).

Criterion	Manual retraining	Cron (7d)	GridSearch (7d)	Method ECLA
Accuracy (t = 100)	0.800	0.830	0.835	0.890
AUC-ROC (t = 100)	0.810	0.840	0.830	0.873
Median TTR, d	14.0	7.0	7.0	1.2
FM (medium)	0.081	0.062	0.058	0.019
No-degradation safeguard	No	No	No	Yes (rollback)
Drift adaptability	No	No	Partial	Yes
Resources/cycle, h	25	25	25+	2–12

The trend in the AUC metric over the course of 10 experimental iterations (Fig. 3) indicates that the proposed Method ECLA exhibits a steady, monotonically increasing performance, whereas the baseline methods reach a plateau as early as iterations (t) 3 and 4.

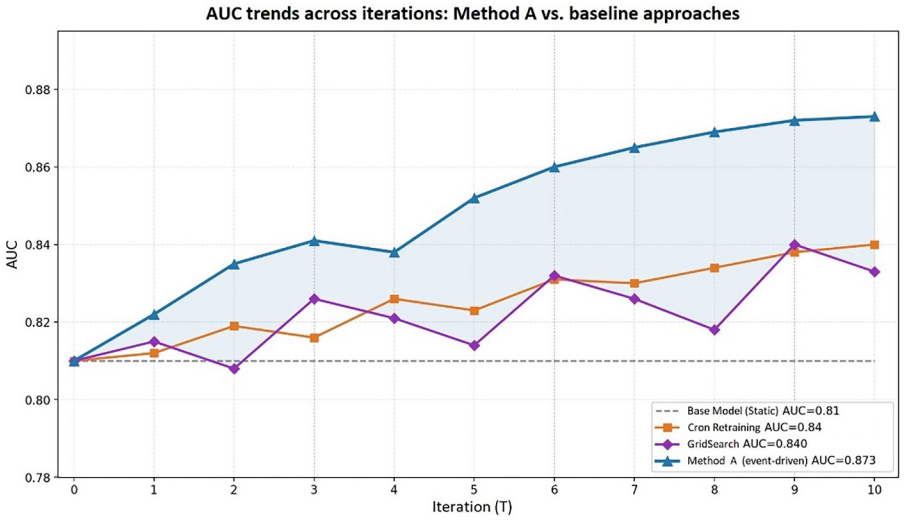

Fig. 3. Changes in the AUC-ROC curve over the course of the iterations.

A comparative analysis of the methods based on classification accuracy and time-to-response (TTR) is presented in Fig. 4.

The statistical significance of the increase in accuracy was verified using the Mann-Whitney U-test. In the early stages of the experiment (at t = 1 and t = 2), the differences in performance were not statistically significant (p = 0.485 and p = 0.285

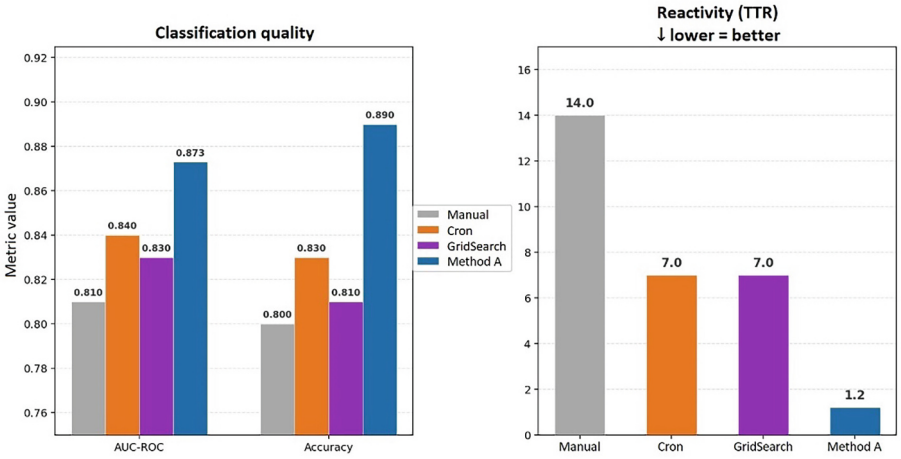

Fig. 4. Comparison of four methods based on AUC-ROC, accuracy, and median TTR in days.

respectively), which is explained by the fact that the volume of accumulated data had not yet reached the threshold value N_{th}. Starting from iteration $t = 5$, the first statistically significant increase in performance was recorded ($p = 0.032$). At $t \geq 10$, a significant difference is observed ($p < 0.001$, $U = 100$), which confirms the complete superiority of Method ECLA: each of the 10 final states of the M_t model demonstrates higher quality metrics compared to all 10 initial states of M_0. The median value of the Accuracy metric shows a monotonic increase with an increment of $\Delta = +0.113$ (in the interval from $t = 1$ to $t = 100$), accompanied by a decrease in variance, indicating increased stability.

The learning curve for Method ECLA, as well as the trend in the AUC-ROC metric for all four methods under consideration (A, B_1, B_2, B_3) over the 180-day experimental period, are shown in Fig. 5.

The learning curve for Method ECLA, plotted based on the results of 10 independent experimental runs, demonstrates steady convergence toward a stable solution. The convergence gap between training accuracy and test accuracy is 0.027 with 20,000 training instances, which is an indication of the absence of overfitting. The $\pm\sigma$ confidence intervals, calculated based on 10 independent experimental runs, show a consistent narrowing as the sample size increases, indicating improved stability and reproducibility of Method ECLA's results as data accumulates.

Another important result of this work is the confirmation of the scalability of Method ECLA. As the data volume increases from 200 to more than 20,000 clinical cases, the local retraining time remains virtually unchanged, since it is determined by the size of the anomaly cluster ($N_{th} \in [33, 50]$) rather than the total database size. This fundamentally distinguishes Method ECLA from basic approaches, for which global retraining time increases linearly with the volume of data: when scaling from 2,000 to 70,000 cases, the duration of manual retraining was 25 h.

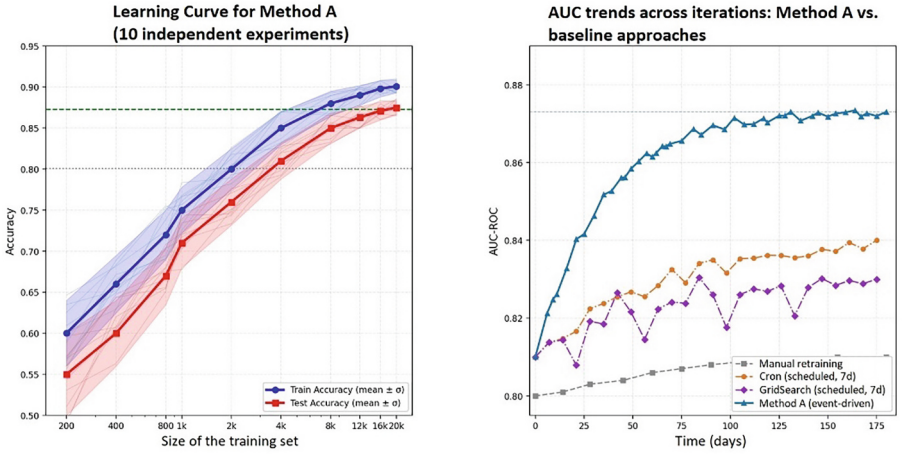

Fig. 5. Learning curve of Method ECLA across 10 independent runs (left panel) and AUC-ROC dynamics for four methods over a 180-day period (right panel).

4 Discussion

The study's findings confirm that the reliable long-term operation of machine learning models in clinical decision support systems requires not only high predictive performance but also robust mechanisms that ensure continuous adaptation to changing data distributions and protection against performance degradation. Traditional retraining strategies typically address only certain aspects of this problem and therefore cannot simultaneously safeguard the model's responsiveness, adaptability, and performance stability.

An important finding of the study is that the superiority of Method ECLA over baseline approaches lies not only in achieving higher final values of the Accuracy and AUC-ROC metrics, as experiments have shown, but also in the stability of performance during the training process.

A significant methodological contribution is the use of a constraint on preserving predictive ability in conjunction with a rollback mechanism $\text{Rollback}(M_t)$. Unlike approaches that rely on implicit assumptions about the reliability of retraining procedures, the proposed method introduces a formal validation criterion for each model update [20–24]. Thus, model adaptation is formalized as a conditional optimization problem with performance preservation safeguard. This constraint is critically important for CDSS, where temporal degradation in prediction quality can directly increase the risk of medical error, thereby jeopardizing the quality of patient care.

A key component of the proposed approach is an anomaly detection procedure based on the Mahalanobis distance, which is particularly applicable to clinical biometric parameters characterized by mutual correlations and heterogeneous measurement scales. A statistically justified threshold derived from the chi-square distribution provides a rigorous mechanism for identifying new clinical patterns that require model adaptation.

While the anomaly detection module assumes multivariate normality, empirical validation confirmed that transformed parameters (log for axial length, keratometry) adhere

to this distribution (Shapiro–Wilk $p > 0.05$). Future ECLA iterations may integrate non-parametric detectors (e.g., Isolation Forest, autoencoders) to mitigate this limitation in heterogeneous populations.

Experimental results demonstrate that integrating incremental learning mechanisms with stability-preserving approaches, including Elastic Weight Consolidation and Experience Replay, enables the system to incorporate new knowledge while preventing catastrophic forgetting. This balance between adaptability and stability is a prerequisite for the sustainable long-term deployment of machine learning systems in medical practice.

An additional advantage of the proposed approach is computational scalability. Since local retraining is triggered by clusters of anomalies rather than the total dataset size, the computational cost of model updates remains stable despite the growing database. This property significantly reduces computational requirements and minimizes the need for manual intervention during system maintenance.

Promising areas for future research include extending the proposed methodology to additional clinical domains and investigating adaptive mechanisms for dynamically optimizing thresholds in anomaly detection procedures.

While the baseline retraining strategies (manual, Cron-based, GridSearch) employed are foundational rather than state-of-the-art in continual learning and concept drift adaptation, this study prioritized validating the feasibility and stability of the proposed event-driven update logic under realistic clinical streaming conditions. Future work will systematically benchmark this approach against advanced methods (e.g., iCaRL, GEM, drift-aware ensembles) to further establish performance boundaries.

References

1. Hinder, F., Vaquet, V., Hammer, B.: One or two things we know about concept drift—a survey on monitoring in evolving environments. Part A: detecting concept drift. Front. Artif. Intell. **7**, 1330257 (2024). https://doi.org/10.3389/frai.2024.1330257
2. Hinder, F., Vaquet, V., Hammer, B.: One or two things we know about concept drift—a survey on monitoring in evolving environments. Part B: locating and explaining concept drift. Front. Artif. Intell. **7**, 1330258 (2024). https://doi.org/10.3389/frai.2024.1330258
3. Garcia, C.M., Abilio, R., Koerich, A.L., Britto, A.S., Barddal, J.P.: Concept drift adaptation in text stream mining settings: a systematic review. ACM Trans. Intell. Syst. Technol. **16**(2), 1–67 (2025). https://doi.org/10.1145/3704922
4. Li, H., Ding, L., Fang, M., Tao, D.: Revisiting catastrophic forgetting in large language model tuning. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-N. (eds.) Findings of ACL: EMNLP 2024, pp. 4297–4308. ACL (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.249
5. Samariya, D., Ma, J., Aryal, S., Zhao, X.: Detection and explanation of anomalies in healthcare data. Health Inf. Sci. Syst. **11**(1), 20 (2023). https://doi.org/10.1007/s13755-023-00221-2
6. McLean, C., Capurro, D.: Concept drift detection to assess the diffusion of process innovations in healthcare. In: AMIA Annual Symposium Proceedings, pp. 746–755 (2023)
7. Kavya, R., et al.: DAAR: drift adaptation and alternatives ranking approach for interpretable clinical decision support systems. Biomed. Signal Process. Control **84**, 104793 (2023). https://doi.org/10.1016/j.bspc.2023.104793
8. Lasko, T.A., Strobl, E.V., Stead, W.W.: Why do probabilistic clinical models fail to transport between sites. npj Digit. Med. **7**(1), 53 (2024). https://doi.org/10.1038/s41746-024-01037-4

9. Chen, Z., et al.: Harnessing the power of clinical decision support systems: challenges and opportunities. Open Heart **10**, e002432 (2023). https://doi.org/10.1136/openhrt-2023-002432
10. da Silva Neto, S.R., et al.: Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: a systematic review. PLoS Negl. Trop. Dis. **16**(1), 1–37 (2022). https://doi.org/10.1371/journal.pntd.0010061
11. Susanto, A.P., Lyell, D., Widyantoro, B., Berkovsky, S., Magrabi, F.: Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. J. Am. Med. Inform. Assoc. **30**(12), 2050–2063 (2023). https://doi.org/10.1093/jamia/ocad180
12. Behrendt, F., et al.: Leveraging the mahalanobis distance to enhance unsupervised brain MRI anomaly detection. In: Linguraru, M.G., et al. (eds.) MICCAI 2024. LNCS, vol. 15011, pp. 394–404. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-72120-5_37
13. Xie, Y., Cruz, L., Heck, P., Rellermeyer, J.S.: Systematic mapping study on the machine learning lifecycle. arXiv preprint arXiv:2103.10248 (2021)
14. Collins, G.S., et al.: TRIPOD+AI statement: Updated guidance for reporting clinical prediction models. BMJ **385**, e078378 (2024). https://doi.org/10.1136/bmj-2023-078378
15. Zozoulenko, N., Cass, T., Gonon, L.: Infinite-dimensional Mahalanobis distance with applications to kernelized novelty detection. arXiv preprint arXiv:2407.11873 (2025)
16. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: theory, method and application. IEEE Trans. Pattern Anal. Mach. Intell. **46**(8), 5362–5383 (2024). https://doi.org/10.1109/TPAMI.2024.3367329
17. Aslam, S., Rasool, A., Li, X., Wu, H.: CEL: a continual learning model for disease outbreak prediction by leveraging domain adaptation via elastic weight consolidation. Interdiscip. Sci. Comput. Life Sci. **17**(2), 390–408 (2025). https://doi.org/10.1007/s12539-024-00675-2
18. Liu, Y., Lian, Z.: PSDNet: A balanced architecture of accuracy and parameters for semantic segmentation. In: ICPR 2020, pp. 827–834 (2021). https://doi.org/10.1109/ICPR48806.2021.9412412
19. Pawar, R., Mukherji, P.: Mann-Whitney U test based identification of significant ECG features for improving accuracy of major depressive disorder classification. In: APCI 2025, pp. 1–6 (2025). https://doi.org/10.1109/APCI65531.2025.11136866
20. Davari, M.R., et al.: Probing representation forgetting in supervised and unsupervised continual learning. arXiv preprint arXiv:2203.13381 (2022)
21. Hassan, S., Rasheed, N., Qureshi, M.A.: A new regularization-based continual learning framework. In: HITE 2024, pp. 1–5 (2024). https://doi.org/10.1109/HITE63532.2024.10777234
22. Gunasekara, N., Pfahringer, B., Gomes, H.M., Bifet, A.: Survey on online streaming continual learning. In: IJCAI 2023, pp. 6628–6637 (2023). https://doi.org/10.24963/ijcai.2023/743
23. van de Ven, G.M., Tuytelaars, T., Tolias, A.S.: Three types of incremental learning. Nat. Mach. Intell. **4**(12), 1185–1197 (2022). https://doi.org/10.1038/s42256-022-00568-3
24. Dalle Pezze, D., Anello, E., Masiero, C., Susto, G.A.: Continual learning approaches for anomaly detection. Evol. Syst. **16**(4), 111 (2025). https://doi.org/10.1007/s12530-025-09732-7