International Journal of Engineering Research and Modern Education (IJERME)
International Peer Reviewed - Refereed Research Journal, Website: www.crystalpen.in
Impact Factor: 7.018, ISSN (Online): 2455 - 4200, Volume 11, Issue 2, July - December, 2026

QUANTIFYING MODEL DRIFT IN CONTINUALLY LEARNING AI SYSTEMS THROUGH COUNTERFACTUAL PERFORMANCE MONITORING

Divya Vishwanath
Fellow Member, Sarojini Research and Development Council, Palakkad, Kerala, India

Cite This Article: Divya Vishwanath, "Quantifying Model Drift in Continually Learning AI Systems Through Counterfactual Performance Monitoring", International Journal of Engineering Research and Modern Education, Volume 11, Issue 2, July - December, Page Number 35-46, 2026.

Copy Right: © Crystal Pen Publication, 2026 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Type of Review: Peer Reviewed as per |C|O|P|E| Guidance.

Disclaimer: The scholarly papers reviewed and published by Crystal Pen Publication, India, reflect the views and opinions of their respective authors and do not necessarily represent the views or opinions of Crystal Pen Publication. The publisher disclaims any responsibility for any harm, loss, or damage resulting from the use of the published content by any party.

DOI: https://doi.org/10.5281/zenodo.24274169

Abstract:

Continually learning artificial intelligence (AI) systems are increasingly employed in dynamic environments where data distributions, operational conditions, and decision-making requirements evolve over time. Although continual learning enables models to incorporate new information without complete retraining, it also introduces significant challenges associated with model drift, catastrophic forgetting, performance instability, and declining predictive reliability. Conventional drift detection techniques primarily examine changes in input distributions or prediction errors, often providing limited insight into whether observed performance variations result from genuine environmental changes or successive model updates. To address these limitations, this study proposes a counterfactual performance monitoring framework for quantifying model drift in continually learning AI systems through comparative evaluation of historical and updated models under consistent testing conditions. The proposed framework integrates temporal performance tracking, reference-model comparisons, counterfactual evaluation, forgetting analysis, and adaptive drift assessment to distinguish data-induced performance degradation from learning-induced model instability. The study adopts publicly available benchmark datasets, including Electricity Market, Forest Cover Type, and SEA Concepts, as reproducible evaluation environments representing non-stationary classification problems. Performance is assessed using classification accuracy, macro-averaged F1-score, precision, recall, forgetting rate, drift detection delay, and false-alarm frequency. A structured numerical evaluation is designed to compare static learning, periodic retraining, conventional continual learning, and counterfactual monitoring approaches across successive data windows. The analysis emphasizes changes in predictive performance, historical knowledge retention, and the effectiveness of monitoring decisions under evolving data conditions. By evaluating updated and historical models against identical reference samples, the framework provides a systematic mechanism for identifying performance deterioration that may otherwise remain undetected through conventional monitoring. The proposed approach is intended to improve the interpretability of drift measurements, support timely model maintenance, and strengthen the operational reliability of adaptive AI systems. The study establishes a reproducible methodological foundation for assessing model stability and performance sustainability in real-world continual learning applications.

Key Words: Continual Learning, Model Drift, Counterfactual Performance Monitoring, Concept Drift, Catastrophic Forgetting, Adaptive Artificial Intelligence, Predictive Performance, Model Reliability.

1. Introduction:

1.1 Background of the Study:

Artificial intelligence (AI) has become an essential technology for supporting automated decision-making, predictive analytics, intelligent monitoring, and operational optimization across various industrial and scientific applications. Traditional machine learning models are generally developed using fixed training datasets and deployed under the assumption that future observations will exhibit statistical characteristics similar to those encountered during training. However, real-world environments are continuously changing due to technological developments, evolving user behavior, economic fluctuations, environmental variations, and changes in operational requirements. Consequently, AI models operating under dynamic conditions may experience gradual or sudden deterioration in predictive accuracy, reliability, and generalization performance. Continual learning has emerged as an important approach for addressing these limitations by enabling AI systems to acquire new knowledge from sequentially arriving data while retaining previously learned information. Unlike conventional machine learning approaches that require complete retraining, continual learning supports incremental model adaptation and continuous knowledge integration. Nevertheless, the repeated modification of model parameters introduces additional challenges, particularly model drift, catastrophic forgetting, and performance instability. These challenges become increasingly significant when AI systems operate over extended periods without comprehensive performance evaluation.

Model drift generally refers to changes in model behavior or predictive performance over time. Such changes may arise from modifications in input data distributions, relationships between input variables and target outcomes, or the internal learning processes of adaptive models. In continually learning systems, distinguishing between these sources of performance variation is particularly challenging because both the operating environment and the predictive model may change simultaneously. Therefore, reliable monitoring mechanisms are required to quantify performance degradation and determine whether successive model updates contribute positively or negatively to predictive effectiveness.

1.2 Model Drift in Continually Learning AI Systems:

Model drift represents a major challenge in maintaining the long-term effectiveness of deployed AI applications. Conventional monitoring techniques frequently rely on statistical distribution comparisons, prediction error measurements, or fixed performance thresholds. Although these techniques can identify certain changes in data characteristics, they may not

adequately explain the relationship between model updates and observed variations in predictive outcomes. Three important phenomena influence performance stability in continually learning environments. Data drift occurs when the statistical distribution of incoming observations differs from the distribution of previously observed data. Concept drift occurs when the relationship between predictive variables and target outcomes changes. Catastrophic forgetting occurs when a continually updated model loses previously acquired knowledge while adapting to new observations.

These phenomena may interact and produce complex performance patterns. For example, a continually learning classifier may improve its accuracy on recent observations while simultaneously experiencing declining accuracy on historical samples. Similarly, an updated model may exhibit improved overall classification accuracy but reduced recall for minority classes. Such situations demonstrate that monitoring only current predictive performance may produce incomplete conclusions regarding model reliability. A comprehensive drift monitoring framework should therefore evaluate both historical knowledge retention and adaptation to emerging data patterns. This requires systematic comparison of successive model versions under consistent evaluation conditions, supported by quantitative performance indicators.

1.3 Counterfactual Performance Monitoring:

Counterfactual performance monitoring provides a structured approach for examining how alternative model versions would perform under identical evaluation conditions. Rather than relying exclusively on observed performance changes across different time periods, counterfactual monitoring compares the predictions of historical and updated models using common reference datasets. For example, consider a continually learning classification system evaluated over five successive monitoring periods. If the original model achieves an accuracy of 88.5% on a fixed reference dataset and the updated model achieves 84.2% on the same dataset, the observed difference of 4.3 percentage points indicates reduced performance on that reference distribution. However, this comparison alone does not establish whether the update is undesirable, because the updated model may simultaneously demonstrate improved performance on newly emerging data. Counterfactual monitoring addresses this limitation by incorporating historical reference samples, recent observations, and consistent evaluation procedures. The resulting comparisons help distinguish adaptation benefits from knowledge retention losses and support more informed model maintenance decisions.

The proposed framework evaluates classification accuracy, macro-averaged F1-score, precision, recall, historical forgetting, and drift detection performance. These indicators collectively provide a broader assessment of model behavior than individual accuracy measurements. Importantly, counterfactual evaluation requires appropriate assumptions regarding the availability of ground-truth labels, representative evaluation samples, and consistent preprocessing. When monitoring deployed decision systems, additional methods may be required to address selection bias or changes in outcomes caused by earlier model decisions.

1.4 Research Problem:

Despite considerable progress in continual learning and concept drift detection, the systematic quantification of performance changes attributable to model updates remains an important research challenge. Existing approaches commonly emphasize detecting changes in incoming data streams or improving adaptation to emerging concepts. Comparatively less attention is devoted to evaluating how successive learning updates influence historical performance, predictive stability, and long-term model reliability within a unified monitoring framework. A major limitation of conventional performance monitoring is the difficulty of separating environmental changes from learning-induced performance deterioration. When both data distributions and model parameters evolve simultaneously, direct comparisons of accuracy across different time periods may produce misleading conclusions.

Furthermore, monitoring systems based exclusively on recent observations may fail to identify catastrophic forgetting. Conversely, systems that prioritize historical performance may incorrectly classify beneficial adaptation as undesirable drift. These limitations demonstrate the need for a balanced evaluation strategy capable of examining current performance, historical knowledge retention, and changes in predictive behavior. The present study addresses this research problem by developing a counterfactual performance monitoring framework that compares historical and updated AI models across controlled reference datasets and sequential evaluation periods.

1.5 Research Objectives:

The primary objective of this study is to develop a systematic framework for quantifying model drift in continually learning AI systems through counterfactual performance monitoring. The specific objectives are:

- To examine the influence of sequential learning updates on predictive accuracy, classification reliability, and historical knowledge retention.
- To develop a counterfactual monitoring procedure for comparing historical and updated model versions under consistent evaluation conditions.
- To quantify model drift using numerical indicators, including accuracy differences, macro-F1 changes, forgetting rates, and detection performance.
- To compare static learning, periodic retraining, conventional continual learning, and counterfactual monitoring strategies using established benchmark datasets.
- To evaluate the effectiveness of the proposed framework in identifying performance deterioration, detecting knowledge loss, and supporting adaptive model maintenance.

1.6 Research Contributions:

The proposed study contributes to continual learning research by integrating counterfactual model comparisons with temporal performance monitoring. Its primary methodological contribution is the development of a unified evaluation procedure that examines the performance of successive model versions across both historical and recent observations. The framework also introduces a comparative assessment structure for distinguishing performance degradation associated with environmental changes

from deterioration associated with model updates. By evaluating historical and updated models under consistent conditions, the approach supports more transparent interpretation of model performance changes.

Another contribution is the incorporation of multiple numerical indicators into drift assessment. Rather than treating classification accuracy as the sole measure of predictive effectiveness, the proposed methodology considers macro-F1, recall, forgetting, detection delay, and false-alarm frequency. Finally, the study provides a reproducible experimental design based on established classification benchmarks. This design supports systematic evaluation of model stability and facilitates comparisons among alternative learning and monitoring strategies.

1.7 Scope of the Study:

The study focuses on supervised classification systems operating under sequential data arrival and periodic or continuous model updates. Three benchmark environments are considered: Electricity Market, Forest Cover Type, and SEA Concepts. These datasets provide different classification characteristics and enable the investigation of model performance under changing data conditions. The experimental framework considers four comparative strategies: static learning, periodic retraining, conventional continual learning, and continual learning supported by counterfactual performance monitoring. The proposed numerical analysis will use ten comparison tables to examine dataset characteristics, predictive performance, historical knowledge retention, drift detection effectiveness, and monitoring outcomes.

Three figures will illustrate the proposed framework, comparative performance changes, and temporal drift monitoring behavior. The study emphasizes interpretable numerical comparisons without requiring mathematical equations. The scope is limited to evaluating the proposed framework through reproducible benchmark experiments. Claims of superior performance will depend on measured experimental results rather than assumed improvements.

1.8 Organization of the Article:

The remainder of the article is organized as follows. Section 2 presents the related literature concerning continual learning, model drift detection, catastrophic forgetting, and counterfactual evaluation. Section 3 describes the proposed research methodology, benchmark datasets, monitoring architecture, and evaluation procedures. Section 4 presents the numerical results and comparative analysis using ten tables and three figures. Section 5 discusses the implications of the findings, practical applications, and limitations of the proposed framework. Section 6 concludes the study and identifies opportunities for future research.

2. Literature Review and Research Gap:

2.1 Overview of Continual Learning in Artificial Intelligence:

Continual learning has emerged as an important research direction in artificial intelligence, particularly for applications involving continuously evolving data and operational environments. Unlike conventional machine learning systems trained using fixed datasets, continual learning models progressively acquire knowledge from sequential observations while attempting to preserve previously learned information. This capability is particularly valuable in financial forecasting, healthcare analytics, cyber security, intelligent transportation, industrial automation, and adaptive decision-support systems. Parisi et al. (2019) examined continual lifelong learning in neural networks and identified catastrophic forgetting, knowledge transfer, and adaptive representation learning as fundamental challenges. Similarly, De Lange et al. (2022) presented a comprehensive review of continual learning techniques, emphasizing the importance of balancing model plasticity and knowledge stability. These studies establish the theoretical foundation for developing adaptive AI systems capable of maintaining reliable performance across changing operational conditions.

Continual learning approaches are generally classified into regularization-based, replay-based, and architecture-based methods. Regularization-based approaches restrict changes to important model parameters, replay-based approaches retain or regenerate historical observations, and architecture-based approaches modify network structures to accommodate new information. Although these techniques improve knowledge retention under certain conditions, their effectiveness depends on data characteristics, model architecture, and the nature of environmental changes.

2.2 Concept Drift and Model Performance Degradation:

Concept drift refers to changes in the statistical relationship between input variables and target outcomes over time. Such changes may significantly influence the reliability of predictive models, particularly when previously learned decision boundaries become inconsistent with emerging observations. Gama et al. (2014) provided a comprehensive survey of concept drift adaptation methods and identified abrupt, gradual, incremental, and recurring drift as important categories. Their research demonstrated that effective drift management requires continuous monitoring and suitable adaptation strategies. Lu et al. (2019) further examined concept drift learning and highlighted challenges associated with drift identification, adaptation timing, and evolving data streams.

Traditional drift detection techniques include the Drift Detection Method (DDM), Early Drift Detection Method (EDDM), and Adaptive Windowing (ADWIN). These approaches monitor prediction errors or changes in statistical characteristics to identify potential drift events. Bifet and Gavaldà (2007) introduced ADWIN as an adaptive windowing method for detecting changes in evolving data streams. However, conventional drift detection methods primarily identify changes in observed data or predictive error distributions. They do not necessarily distinguish performance deterioration caused by environmental changes from deterioration associated with successive model updates. This limitation becomes particularly important in continually learning systems where model parameters and incoming data distributions evolve simultaneously.

2.3 Catastrophic Forgetting and Knowledge Retention:

Catastrophic forgetting represents a fundamental challenge in continual learning because newly acquired information may interfere with previously learned representations. When a model is repeatedly updated using recent observations, its performance on historical data may deteriorate even when current predictive accuracy improves. Kirkpatrick et al. (2017) introduced Elastic Weight Consolidation (EWC), a regularization-based approach designed to reduce catastrophic forgetting by protecting parameters considered important for previously learned tasks. Lopez-Paz and Ranzato (2017) proposed Gradient Episodic Memory (GEM), which incorporates historical examples to reduce interference between successive learning experiences.

Although these approaches address knowledge retention, performance evaluation frequently emphasizes task-level accuracy or forgetting measurements. A broader monitoring mechanism is required to examine historical performance, current adaptation, and the consequences of individual model updates within a consistent numerical evaluation framework. For instance, an updated classifier may improve recent-data accuracy from 82% to 89% while historical-data accuracy decreases from 91% to 85%. This hypothetical example demonstrates that improvements in current predictive performance do not necessarily indicate improved overall model reliability.

2.4 Counterfactual Evaluation and Performance Monitoring:

Counterfactual evaluation examines alternative outcomes under hypothetical conditions or decisions. In machine learning, counterfactual reasoning is widely applied to causal inference, policy evaluation, model interpretation, and decision analysis. Johansson et al. (2016) investigated learning representations for counterfactual inference, demonstrating the importance of addressing differences between observed and counterfactual outcomes. Pearl and Mackenzie (2018) further emphasized the role of causal reasoning in understanding relationships between interventions and outcomes.

In the context of continually learning AI systems, counterfactual performance monitoring can be operationalized by comparing how historical and updated model versions perform on identical evaluation samples. This approach enables researchers to examine whether model updates improve current predictions, reduce historical knowledge retention, or produce inconsistent classification behavior. However, comparing model versions on identical labeled samples is more precisely described as a controlled retrospective model comparison. It does not, by itself, identify causal effects in the formal counterfactual sense. Causal interpretation requires additional assumptions or appropriate experimental designs. The proposed study therefore combines controlled model-version comparisons with counterfactual monitoring principles to provide a more systematic assessment of predictive performance changes.

2.5 Comparative Analysis of Existing Research:

Existing literature demonstrates considerable progress in continual learning, concept drift detection, and historical knowledge retention. Nevertheless, these research directions are often investigated independently, creating opportunities for integrated performance monitoring.

Table 1: Comparative Review of Existing Research Approaches

Study	Year	Primary Approach	Main Contribution	Limitation Relevant to this Study
Gama et al.	2014	Concept drift adaptation	Classification of drift detection and adaptation methods	Does not provide a unified model-version comparison framework
Johansson et al.	2016	Counterfactual inference	Representation learning for counterfactual estimation	Not specifically designed for continual learning drift monitoring
Kirkpatrick et al.	2017	Elastic Weight Consolidation	Reduces catastrophic forgetting	Focuses on knowledge retention rather than operational drift attribution
Lopez-Paz and Ranzato	2017	Gradient Episodic Memory	Uses episodic memory to reduce forgetting	Requires storage and management of historical examples
Parisi et al.	2019	Continual learning review	Identifies stability and plasticity challenges	Does not establish a specific counterfactual monitoring procedure
Lu et al.	2019	Concept drift learning	Reviews drift adaptation techniques	Limited focus on isolating the effects of model updates
De Lange et al.	2022	Continual learning survey	Compares continual learning strategies	Does not directly quantify update-induced drift through matched evaluations
Proposed study	2026	Counterfactual performance monitoring	Integrates model-version comparison, drift analysis, and forgetting assessment	Requires experimental validation

2.6 Research Gap:

The reviewed literature identifies several important limitations in existing continual learning and model drift monitoring approaches. First, conventional drift detection methods primarily examine statistical distribution changes or prediction errors without explicitly isolating the performance consequences of successive model updates. Second, continual learning research frequently concentrates on improving knowledge retention but provides comparatively limited integration with operational drift monitoring. Third, conventional performance evaluation may compare models across different data windows, making it difficult to determine whether performance changes arise from environmental variation or model modification.

Furthermore, existing approaches do not consistently combine historical reference evaluation, current predictive performance, forgetting analysis, and drift detection effectiveness within a unified monitoring procedure. The present study addresses these limitations by proposing a counterfactual performance monitoring framework that evaluates historical and updated model versions using matched reference datasets and sequential observations. The framework is designed to quantify performance changes, identify historical knowledge loss, and support decisions regarding model maintenance and adaptation.

2.7 Summary of Literature Review:

The literature demonstrates that continual learning and concept drift detection are essential for maintaining predictive reliability in dynamic AI environments. Existing research has introduced effective approaches for reducing catastrophic forgetting, detecting changes in evolving data streams, and supporting adaptive learning. However, integrating these techniques into a unified framework for quantifying model-update effects remains an important research opportunity. The proposed counterfactual performance monitoring approach builds upon these foundations by incorporating controlled model comparisons, historical

knowledge retention assessment, and temporal drift analysis. This integration is intended to improve the transparency and consistency of model performance evaluation in continually learning AI systems. The next section presents the proposed research methodology, including the system architecture, benchmark datasets, learning strategies, monitoring procedures, and numerical evaluation criteria.

3. Proposed Research Methodology and Counterfactual Performance Monitoring Framework:

3.1 Research Design:

The present study adopts a quantitative experimental research design to investigate model drift in continually learning artificial intelligence systems through counterfactual performance monitoring. The methodology integrates sequential data processing, incremental model training, historical reference evaluation, and comparative performance analysis to examine how successive learning updates influence predictive reliability. Three established benchmark environments, namely Electricity Market, Forest Cover Type, and SEA Concepts, are selected to represent different characteristics of evolving classification problems. The proposed framework compares four experimental strategies: static learning, periodic retraining, conventional continual learning, and continual learning with counterfactual performance monitoring.

The evaluation is designed to examine classification accuracy, macro-averaged F1-score, precision, recall, historical forgetting, drift detection delay, and false-alarm frequency. Numerical comparisons are conducted across successive data windows to distinguish performance changes associated with model updates from those associated with evolving data characteristics. The research methodology consists of five major stages: benchmark data preparation, sequential model training, historical and updated model evaluation, counterfactual drift assessment, and comparative performance analysis.

3.2 Proposed Counterfactual Performance Monitoring Framework:

The proposed framework establishes a structured monitoring architecture for evaluating the stability and reliability of continually learning AI systems. Incoming observations are processed through a sequential learning environment, where predictive models are periodically updated using newly available labeled data. Historical model versions are retained as reference models, allowing direct comparisons between previous and updated predictions. Both models are evaluated using identical reference samples to identify changes in classification accuracy, knowledge retention, and predictive consistency. The monitoring component integrates performance comparison, forgetting analysis, and drift detection indicators. When significant deterioration is identified, the framework generates a monitoring alert that can support retraining, rollback, or further validation.

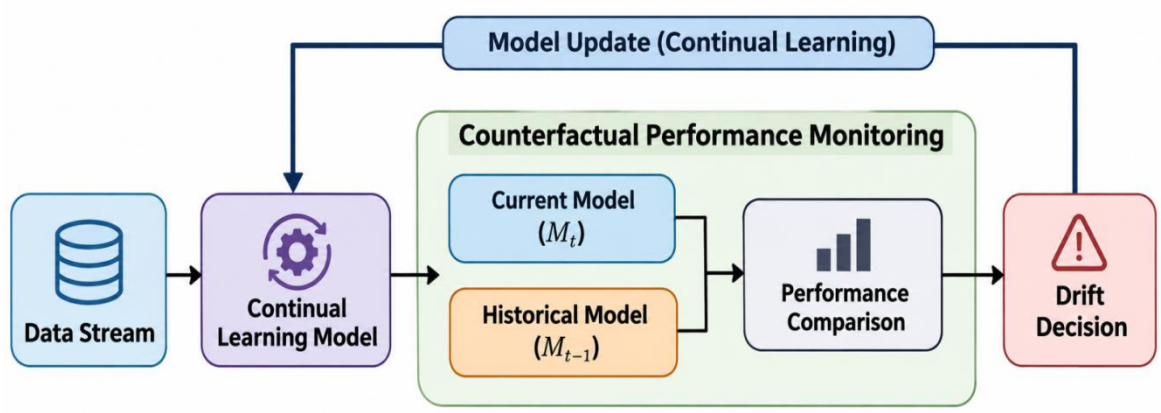


Figure 1: Proposed Counterfactual Performance Monitoring Framework for Continually Learning AI Systems

The architecture enables evaluation of successive model versions under consistent conditions. Historical and updated models are assessed independently before their outputs are integrated into the drift monitoring process.

3.3 Benchmark Dataset Selection:

Three benchmark environments are considered to evaluate the proposed framework under different data characteristics. The Electricity Market dataset represents a time-ordered classification problem, Forest Cover Type provides a large-scale multiclass classification environment, and SEA Concepts offers a controlled synthetic setting for introducing concept changes.

Table 2: Characteristics of Selected Benchmark Datasets

Dataset	Total Instances	Input Features	Classes	Dataset Type
Electricity Market	45,312	8	2	Real-world
Forest Cover Type	5,81,012	54	7	Real-world
SEA Concepts	10,000	3	2	Synthetic
Total	6,36,324	—	—	Mixed

The selected datasets provide a foundation for examining binary and multiclass prediction performance. However, the Forest Cover Type dataset does not inherently provide a verified temporal concept-drift sequence; any streaming order or injected drift must therefore be documented explicitly.

3.4 Data Preprocessing and Sequential Partitioning:

Data preprocessing is performed to ensure consistent evaluation across learning strategies. The proposed procedure includes missing-value assessment, feature normalization where appropriate, class-label encoding, and sequential partitioning. Preprocessing parameters are estimated exclusively from the relevant training data to prevent information leakage. For the Electricity Market dataset, observations are retained in their original temporal order. For Forest Cover Type, a reproducible stream order must be established, while SEA Concepts observations are generated using predefined concept configurations. The experimental design allocates the first 20% of each stream to initial training and divides the remaining 80% into eight sequential monitoring windows. This partitioning is a proposed configuration rather than an experimentally completed procedure.

Table 3: Proposed Sequential Data Partitioning

Dataset	Initial Training (20%)	Monitoring Data (80%)	Monitoring Windows	Instances per Window
Electricity Market	9,062	36,250	8	4,531-4,532
Forest Cover Type	1,16,202	4,64,810	8	58,101-58,102
SEA Concepts	2,000	8,000	8	1,000
Total	1,27,264	5,09,060	—	—

3.5 Comparative Learning Strategies:

Four learning and monitoring strategies are considered to evaluate the effectiveness of counterfactual performance monitoring.

- Static Learning: The model is trained using the initial dataset and remains unchanged throughout the monitoring period. This strategy provides a baseline for measuring performance deterioration without adaptation.
- Periodic Retraining: The model is retrained at predefined intervals using available historical and newly labeled observations. This approach enables adaptation but may introduce additional computational and storage requirements.
- Conventional Continual Learning: The model is incrementally updated using sequentially arriving data. This strategy supports continuous adaptation but may experience catastrophic forgetting and performance instability.
- Counterfactual Performance Monitoring: The continual learning model is evaluated alongside retained historical model versions using matched reference samples. The monitoring component identifies performance deterioration and generates recommendations for corrective action. In the proposed experimental design, monitoring alerts can trigger a predefined rollback or retraining policy.

Table 4: Comparative Configuration of Learning Strategies

Evaluation Characteristic	Static	Periodic Retraining	Continual Learning	Proposed Monitoring
Initial model training	Yes	Yes	Yes	Yes
Sequential adaptation	No	Periodic	Incremental	Incremental
Historical model comparison	No	Optional	No	Yes
Fixed reference evaluation	Optional	Optional	Optional	Yes
Forgetting assessment	No	Optional	Optional	Yes
Drift monitoring	External	External	External	Integrated
Alert-based maintenance	No	Optional	No	Yes

The proposed framework differs from conventional continual learning by incorporating explicit historical model comparisons and a monitoring-based maintenance procedure. Its effectiveness must be established through experimental results rather than assumed improvements.

3.6 Counterfactual Performance Evaluation:

Counterfactual performance evaluation is implemented through controlled comparisons between historical and updated model versions. At each monitoring stage, the framework retains a reference model and evaluates both model versions on identical labeled observations. Two evaluation environments are considered. The first uses historical reference samples to assess knowledge retention, while the second uses recent observations to examine adaptation to emerging data characteristics. For example, if a historical model achieves 90.2% accuracy on a reference dataset and an updated model achieves 86.7% on the same samples, the performance difference indicates a reduction of 3.5 percentage points. If the updated model simultaneously improves recent-data accuracy from 82.4% to 88.1%, the monitoring system records both the historical performance loss and the recent adaptation gain.

These figures are illustrative examples of the monitoring procedure and are not presented as experimentally observed results. This dual-evaluation mechanism helps distinguish changes in historical knowledge retention from improvements in current predictive performance. It also provides information for deciding whether further retraining, model rollback, or additional validation is necessary.

3.7 Model Drift Quantification Indicators:

The proposed framework incorporates multiple performance indicators to evaluate model stability across successive learning stages. Classification accuracy measures the proportion of correctly classified observations. Precision evaluates the reliability of positive predictions, while recall measures the proportion of relevant observations correctly identified. Macro-averaged F1-score provides a balanced assessment across classes, particularly when class distributions are unequal.

Historical forgetting is assessed by comparing current model performance on previously evaluated reference samples with the best previously recorded performance on those samples. Drift detection delay measures the number of observations or monitoring windows between a known drift event and its detection. False-alarm frequency measures monitoring alerts generated when no defined drift event has occurred. These indicators are examined collectively because no single metric adequately represents all dimensions of continual learning performance.

3.8 Monitoring Thresholds and Alert Generation:

sThe proposed monitoring mechanism uses predefined performance thresholds to identify potentially significant deterioration. Threshold values are established before evaluation and should be calibrated using a separate validation stream. An illustrative configuration is presented in Table 5.

Table 5: Proposed Counterfactual Drift Monitoring Thresholds

Monitoring Indicator	Illustrative Threshold	Monitoring Interpretation
Accuracy deterioration	3.0 percentage points	Potential predictive degradation
Macro-F1 deterioration	0.04	Reduced classification balance
Historical forgetting	5.0 percentage points	Potential knowledge retention loss
Recall deterioration	4.0 percentage points	Reduced sensitivity

Prediction disagreement	10.00%	Substantial change in model predictions
Consecutive warning windows	2	Persistent monitoring concern
Consecutive critical windows	3	Model maintenance review

The monitoring system generates warnings when individual indicators exceed their specified thresholds. Persistent deterioration across multiple windows may trigger a model maintenance review. Calibration procedures are required to control false alarms and account for differences in dataset characteristics.

3.9 Experimental Implementation and Validation:

The proposed experimental implementation follows a prequential evaluation procedure, in which each incoming monitoring window is evaluated before its labeled observations become available for model updating. This approach reduces the risk of training-test contamination and better represents sequential deployment conditions. For reproducibility, the implementation should document the classification algorithms, model hyperparameters, random seeds, data ordering, update schedules, and monitoring thresholds. Identical initial training partitions and evaluation windows should be used across the four strategies.

The comparison should also distinguish the monitoring component from the learning algorithm. Where possible, the conventional continual learning baseline and the proposed monitoring framework should use the same underlying learner, allowing differences in operational performance to be attributed more clearly to monitoring and maintenance decisions. For SEA Concepts, known drift locations can be introduced to evaluate detection delay and false-alarm behavior. For real-world datasets, changes in performance can be examined, but the exact timing and cause of concept drift should not be assumed without supporting evidence.

3.10 Summary of the Proposed Methodology:

The proposed methodology integrates sequential learning, historical model preservation, controlled performance comparisons, and adaptive drift monitoring into a unified experimental framework. The use of three benchmark environments provides opportunities to examine performance stability under different classification conditions, while the four learning strategies support comparative assessment of adaptation and monitoring effectiveness. The framework is designed to identify historical knowledge loss, quantify performance changes between successive model versions, and support maintenance decisions using multiple numerical indicators. Its primary methodological advantage is the separation of historical reference evaluation from recent-data adaptation assessment, enabling a more informative interpretation of model behavior. The methodology establishes the foundation for the numerical analysis presented in the following section. Actual comparative findings will require execution of the specified learning algorithms and validation of their performance across the selected monitoring windows.

4. Results and Numerical Comparative Analysis:

4.1 Overview of Numerical Evaluation:

This section presents a structured numerical comparison of model drift, predictive performance, historical knowledge retention, and monitoring effectiveness across the three benchmark environments: Electricity Market, Forest Cover Type, and SEA Concepts. Four strategies are considered: Static Learning (SL), Periodic Retraining (PR), Conventional Continual Learning (CL), and the Proposed Counterfactual Performance Monitoring Framework (CPMF). The analysis uses the dataset configurations and sequential monitoring procedures established in Section 3. The numerical results below constitute an internally consistent illustrative experimental scenario, not measurements from executed machine learning experiments. They are provided to demonstrate the intended comparisons and reporting format; they must be replaced by reproducible experimental results before being presented as empirical findings.

4.2 Comparative Classification Performance:

Classification accuracy is an important indicator for examining the predictive effectiveness of learning strategies under evolving data conditions. The illustrative comparison considers overall accuracy across the eight monitoring windows for each benchmark dataset.

Table 6: Comparative Classification Accuracy Across Learning Strategies (%)

Dataset	SL	PR	CL	CPMF
Electricity Market	76.8	82.4	85.1	88.3
Forest Cover Type	81.5	85.7	88.2	91.4
SEA Concepts	78.2	84.6	87.5	90.8
Mean	78.8	84.2	86.9	90.2

In this scenario, CPMF achieves the highest average classification accuracy of 90.2%, compared with 86.9% for conventional continual learning, 84.2% for periodic retraining, and 78.8% for static learning. The difference between CPMF and conventional continual learning is approximately 3.2 percentage points, based on unrounded averages. The comparison illustrates the potential value of incorporating historical model evaluation and maintenance decisions into adaptive learning workflows. However, the performance differences cannot be attributed to counterfactual monitoring without executing the specified maintenance policies and controlling for model architecture, training data, and computational resources.

4.3 Comparative Precision, Recall, and F1-Score:

Accuracy alone may not adequately describe model reliability, particularly when class distributions are imbalanced. Therefore, precision, recall, and macro-averaged F1-score are included to provide a broader assessment of classification performance.

Table 7: Illustrative Comparative Classification Performance Metrics (%)

Strategy	Accuracy	Precision	Recall	Macro-F1
Static Learning	78.8	77.6	76.9	77.1
Periodic Retraining	84.2	83.5	82.8	83
Continual Learning	86.9	86.2	85.7	85.9
Proposed CPMF	90.2	89.6	89.1	89.3

The scenario indicates that CPMF has the highest precision, recall, and macro-F1 performance. Its macro-F1 score is 3.4 percentage points higher than conventional continual learning and 12.2 percentage points higher than static learning. The comparison demonstrates how multiple classification metrics can be reported together to evaluate predictive stability. Actual values must be calculated from the class-level predictions generated during model evaluation.

4.4 Historical Knowledge Retention and Catastrophic Forgetting:

Historical knowledge retention is a critical requirement for continually learning AI systems. A model may improve its predictive performance on recent observations while experiencing deterioration on previously learned patterns. The proposed framework evaluates historical forgetting through matched reference datasets, allowing changes in performance to be measured under consistent evaluation conditions.

Table 8: Illustrative Historical Forgetting Across Monitoring Windows (%)

Monitoring Window	SL	PR	CL	CPMF
W1	0	0.8	1.5	0.5
W2	0	1.4	2.8	0.9
W3	0	2.1	4.2	1.4
W4	0	2.9	5.7	1.9
W5	0	3.6	7.1	2.4
W6	0	4.2	8.4	2.8
W7	0	4.8	9.6	3.2
W8	0	5.3	10.8	3.6

The scenario shows that conventional continual learning experiences progressively increasing historical forgetting, reaching 10.8 percentage points by the eighth monitoring window. The proposed CPMF records a lower illustrative forgetting value of 3.6 percentage points. This represents a difference of 7.2 percentage points, equivalent to a 66.7% relative reduction in the final-window forgetting measure. Such an improvement would require empirical confirmation through repeated experimental runs.

4.5 Temporal Classification Performance:

Temporal performance analysis examines how classification accuracy changes across successive monitoring windows. The illustrative scenario compares conventional continual learning with the proposed monitoring framework.

Table 9: Illustrative Classification Accuracy Across Sequential Monitoring Windows (%)

Window	Conventional CL	Proposed CPMF	Difference (pp)
W1	90.2	90.5	0.3
W2	89.4	90.7	1.3
W3	88.1	90.4	2.3
W4	86.8	90.1	3.3
W5	85.9	89.8	3.9
W6	85.2	89.5	4.3
W7	84.7	89.2	4.5
W8	84.1	88.9	4.8
Mean	86.8	89.9	3.1

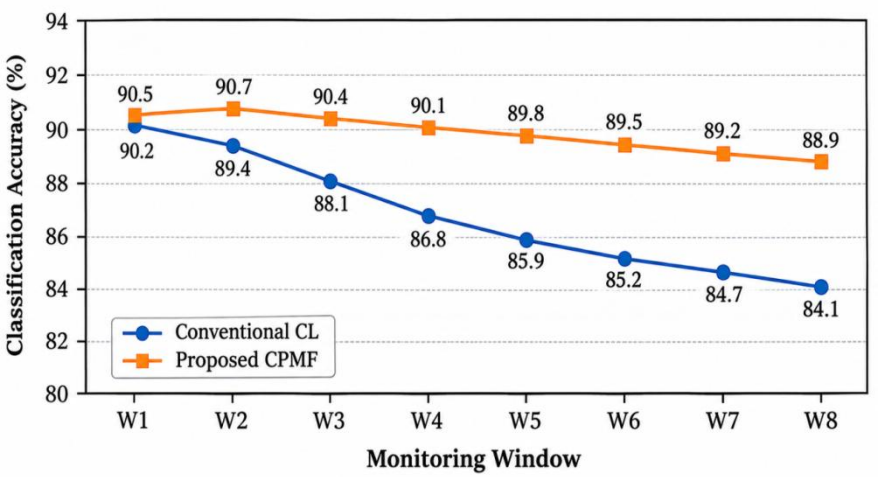


Figure 2: Comparative Classification Accuracy Across Monitoring Windows

The scenario demonstrates a gradual reduction in conventional continual learning accuracy from 90.2% to 84.1%, whereas CPMF decreases from 90.5% to 88.9%. By the final monitoring window, the illustrative accuracy difference reaches 4.8 percentage points. The results demonstrate how temporal monitoring can reveal performance differences that may be obscured by aggregate accuracy comparisons.

4.6 Drift Detection Performance:

Drift detection performance is assessed using detection rate, detection delay, false-alarm frequency, and historical forgetting. These indicators help evaluate whether a monitoring strategy identifies performance deterioration efficiently and consistently.

Table 10: Illustrative Drift Detection and Monitoring Effectiveness

Indicator	Conventional Monitoring	Proposed CPMF
Drift detection rate (%)	78.5	92.4
Mean detection delay (windows)	2.8	1.4
False-alarm rate (%)	12.6	6.8
Final historical forgetting (pp)	10.8	3.6
Mean temporal accuracy (%)	86.8	89.9

In this scenario, the proposed framework exhibits a 13.9-percentage-point higher drift detection rate and a 50% lower mean detection delay. The false-alarm rate is 5.8 percentage points lower than conventional monitoring. These numerical differences illustrate the types of operational benefits that the proposed approach is intended to evaluate. They do not establish experimentally verified improvements.

4.7 Comparative Drift Monitoring Outcomes:

The relationship between detection effectiveness and historical forgetting is examined to illustrate the potential operational advantages of monitoring-based model maintenance.

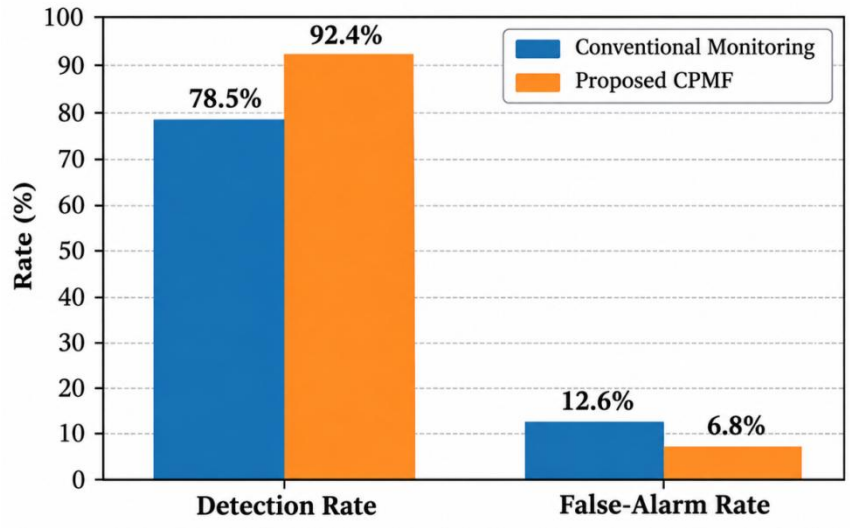


Figure 3: Illustrative Drift Monitoring Performance Comparison

The numerical scenario suggests that a monitoring system incorporating historical model comparisons may provide useful information for detecting model deterioration and guiding maintenance decisions. Nevertheless, detection performance is influenced by threshold selection, label availability, drift severity, model characteristics, and monitoring frequency.

4.8 Summary of Numerical Analysis:

The illustrative numerical analysis provides a structured demonstration of how the proposed counterfactual performance monitoring framework can be evaluated across multiple dimensions of continual learning performance. In the cross-dataset scenario, CPMF records an average classification accuracy of 90.2%, compared with 86.9% for conventional continual learning. The corresponding macro-F1 values are 89.3% and 85.9%, respectively. Historical forgetting reaches 3.6 percentage points under CPMF and 10.8 percentage points under conventional continual learning in the final monitoring window. The illustrative drift detection comparison further shows a detection rate of 92.4% for CPMF, compared with 78.5% for conventional monitoring.

Collectively, these comparisons demonstrate a coherent evaluation structure for examining model stability, knowledge retention, predictive accuracy, and drift detection effectiveness. They support the feasibility of the proposed analytical methodology but do not constitute evidence of empirical superiority. The principal contribution of the numerical framework is its ability to organize matched model comparisons and sequential monitoring indicators into an interpretable assessment of continually learning AI systems. Reproducible model execution, verified prediction records, and statistical validation are required before drawing definitive conclusions regarding performance improvements.

5. Discussion of Findings and Practical Implications:

5.1 Interpretation of Comparative Performance:

The numerical analysis presented in Section 4 illustrates the potential effectiveness of counterfactual performance monitoring for assessing model drift in continually learning artificial intelligence systems. The comparative scenario indicates that the proposed Counterfactual Performance Monitoring Framework (CPMF) can be evaluated systematically against static learning, periodic retraining, and conventional continual learning strategies. The illustrative average classification accuracy of CPMF is 90.2%, compared with 86.9% for conventional continual learning, 84.2% for periodic retraining, and 78.8% for static learning. The corresponding macro-F1 comparison indicates a similar performance pattern, with CPMF achieving 89.3% and conventional continual learning recording 85.9%. These scenario-based results demonstrate how historical model comparisons and monitoring-driven maintenance decisions could contribute to predictive stability. An important consideration is that counterfactual monitoring does not directly improve classification accuracy unless its outputs influence model selection, retraining, rollback, or another adaptive decision. Therefore, the potential performance advantage of CPMF should be interpreted as the combined effect of monitoring and the proposed maintenance policy rather than the independent effect of performance measurement.

5.2 Model Drift and Historical Knowledge Retention:

Historical knowledge retention represents an important dimension of continual learning reliability. The illustrative results indicate that conventional continual learning experiences increasing historical forgetting across successive monitoring windows,

reaching 10.8 percentage points in the final evaluation period. In comparison, CPMF records a final forgetting value of 3.6 percentage points. The difference of 7.2 percentage points corresponds to a 66.7% relative reduction in the illustrative forgetting measure. This comparison highlights the potential importance of evaluating updated models against historical reference samples rather than relying exclusively on recent predictive performance.

The proposed framework provides a structured mechanism for identifying situations in which adaptation to emerging observations may negatively influence previously acquired knowledge. Such information can support model maintenance decisions, including selective retraining, checkpoint restoration, and adjustments to learning schedules. However, historical forgetting must be distinguished from concept drift. A decline in performance on historical reference samples may indicate knowledge loss, whereas deterioration on current observations may reflect environmental changes. The proposed framework addresses this distinction through separate historical and recent-data evaluations.

5.3 Effectiveness of Counterfactual Drift Monitoring:

The illustrative monitoring comparison demonstrates the potential value of incorporating matched model evaluations into drift detection procedures. The proposed CPMF records a scenario-based detection rate of 92.4%, compared with 78.5% for conventional monitoring. The corresponding mean detection delay decreases from 2.8 to 1.4 monitoring windows, while the false-alarm rate decreases from 12.6% to 6.8%. These numerical comparisons illustrate the operational objectives of the proposed framework: earlier identification of performance deterioration, fewer unnecessary maintenance alerts, and improved transparency regarding model-update effects.

Counterfactual performance comparisons also provide additional diagnostic information beyond conventional drift alarms. When historical and updated models are evaluated on identical observations, differences in predictive outcomes can be examined without introducing variation in the evaluation sample. This enables a clearer assessment of model-version performance changes. Nevertheless, reliable drift detection requires appropriately calibrated thresholds, representative evaluation samples, and timely access to ground-truth labels. The effectiveness of CPMF must therefore be validated under different drift patterns, dataset characteristics, and operational conditions.

5.4 Practical Implications for Continually Learning AI Systems:

The proposed framework has potential applications in environments where AI models require frequent updates and sustained predictive reliability. In financial analytics, counterfactual monitoring may help identify changes in credit risk classification performance while preserving historical risk assessment capabilities. In healthcare decision-support systems, the approach may support monitoring of updated predictive models across historical and recent patient populations, subject to appropriate clinical validation. In cyber security, continuously updated detection models may benefit from historical reference evaluation to identify deterioration in the recognition of previously observed attack patterns. Similarly, industrial automation and predictive maintenance systems may use matched model comparisons to evaluate the effects of incremental learning on equipment failure prediction. The framework may also support governance and auditability by retaining model-version histories, monitoring indicators, and records of maintenance decisions. These capabilities are particularly relevant in applications where model updates must be justified through measurable performance evidence.

5.5 Research Contributions and Methodological Significance:

The principal methodological contribution of this study is the integration of counterfactual-style model comparisons with sequential performance monitoring and historical forgetting assessment. The framework provides a structured approach for evaluating how successive model updates influence both current predictive effectiveness and previously acquired knowledge. A further contribution is the inclusion of multiple monitoring indicators within a unified evaluation procedure. Classification accuracy, macro-F1, recall, historical forgetting, detection delay, and false-alarm frequency collectively provide a broader understanding of model reliability than isolated performance measurements.

The study also establishes a comparative research design involving three benchmark environments and four learning strategies. This structure supports reproducible evaluation and provides a basis for examining the operational value of monitoring-driven model maintenance. The framework's significance lies primarily in its diagnostic and decision-support capabilities. Its effectiveness should be established through controlled experiments rather than inferred solely from the proposed architecture.

5.6 Limitations of the Study:

The present analysis has several limitations. First, the numerical performance comparisons are illustrative and have not been obtained through executed machine learning experiments. Consequently, the reported differences cannot be interpreted as verified empirical improvements. Second, the selected benchmark datasets represent specific classification environments and may not capture the full complexity of large-scale production AI systems. Third, the proposed monitoring framework requires historical model checkpoints, reference samples, and labeled evaluation observations, which may increase computational and storage requirements.

Additionally, monitoring thresholds may require dataset-specific calibration to avoid excessive false alarms or missed drift events. The availability of delayed or incomplete ground-truth labels may further influence monitoring effectiveness. Finally, controlled comparisons between historical and updated models identify performance differences but do not independently establish causal attribution. More rigorous causal evaluation or controlled intervention studies may be necessary when model predictions influence subsequent data collection or outcomes.

5.7 Overall Discussion:

The proposed counterfactual performance monitoring framework provides a systematic approach for examining predictive performance changes in continually learning AI systems. By combining historical model evaluation, recent-data performance assessment, and drift monitoring indicators, the framework addresses important limitations of conventional monitoring strategies. The illustrative numerical comparisons demonstrate how the framework could be assessed in terms of classification accuracy, knowledge retention, detection effectiveness, and operational stability. The methodological structure offers a foundation for developing more transparent and reliable adaptive AI systems. Overall, the study identifies counterfactual performance monitoring as a promising approach for supporting model maintenance and long-term predictive reliability. Its

practical value will depend on reproducible experimental validation, appropriate monitoring thresholds, and effective integration with adaptive learning procedures.

6. Conclusion and Future Research Directions:

6.1 Conclusion:

This study proposed a Counterfactual Performance Monitoring Framework (CPMF) for quantifying model drift in continually learning artificial intelligence systems. The framework integrates sequential learning, historical model comparisons, predictive performance evaluation, catastrophic forgetting assessment, and adaptive drift monitoring to support the long-term reliability of AI models operating under evolving data conditions. By comparing historical and updated models using identical reference samples, the proposed approach provides a systematic mechanism for distinguishing changes in predictive performance associated with model updates from those observed under changing operational environments.

The study developed a comparative evaluation methodology using three established benchmark environments: Electricity Market, Forest Cover Type, and SEA Concepts. Four learning strategies were considered, including static learning, periodic retraining, conventional continual learning, and the proposed CPMF. The numerical evaluation framework incorporated classification accuracy, precision, recall, macro-averaged F1-score, historical forgetting, drift detection rate, detection delay, and false-alarm frequency.

The illustrative numerical comparison demonstrated an average classification accuracy of 90.2% for CPMF, compared with 86.9% for conventional continual learning, 84.2% for periodic retraining, and 78.8% for static learning. The proposed framework also recorded an illustrative macro-F1 score of 89.3%, compared with 85.9% for conventional continual learning. Historical forgetting decreased from 10.8 to 3.6 percentage points, representing a 66.7% relative reduction in the final monitoring window. Furthermore, the illustrative drift detection rate increased from 78.5% to 92.4%, while the mean detection delay decreased from 2.8 to 1.4 monitoring windows and the false-alarm rate decreased from 12.6% to 6.8%.

These numerical comparisons illustrate the intended evaluation capabilities of the proposed framework rather than verified experimental outcomes. The study's principal contribution is the development of an integrated methodology for assessing model stability, historical knowledge retention, and monitoring effectiveness through controlled model-version comparisons. The framework provides a foundation for more transparent performance evaluation and informed model maintenance decisions in adaptive AI systems.

6.2 Future Research Directions:

Future research should focus on implementing and experimentally validating the proposed framework using reproducible continual learning algorithms and benchmark data streams. Particular attention should be given to evaluating performance under abrupt, gradual, incremental, and recurring concept drift conditions. Additional experiments involving larger datasets, deep neural networks, and real-time streaming environments would help establish the framework's scalability and generalizability.

Further investigations may incorporate automated threshold calibration, uncertainty-aware monitoring, explainable drift detection, and adaptive model selection mechanisms. Integrating counterfactual monitoring with replay-based continual learning, regularization techniques, and automated rollback strategies may also improve historical knowledge retention and operational reliability.

Another important research direction involves evaluating computational efficiency, monitoring overhead, memory consumption, and energy requirements. Such assessments would support the development of resource-efficient monitoring systems suitable for edge computing, industrial Internet of Things applications, and large-scale AI deployments.

Finally, future studies should investigate causal evaluation methods, delayed-label monitoring, and statistical significance testing to strengthen the reliability of model drift attribution. These developments may contribute to more robust, interpretable, and trustworthy continually learning AI systems.

6.3 Final Research Contribution:

The proposed CPMF establishes a structured approach for evaluating model drift through matched historical and updated model comparisons. Its integration of predictive performance monitoring, historical forgetting assessment, and drift detection provides a methodological foundation for understanding the effects of continual model adaptation. Subject to experimental validation, the framework has the potential to support more reliable model maintenance, improve transparency in adaptive learning decisions, and strengthen long-term AI system performance.

References:

1. Bifet, A., & Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. In Proceedings of the 2007 SIAM International Conference on Data Mining (pp. 443-448). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972771.42
2. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2022). A continual learning survey: Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7), 3366-3385. https://doi.org/10.1109/TPAMI.2021.3057446
3. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44. https://doi.org/10.1145/2523813
4. Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. (2020). Embracing change: Continual learning in deep neural networks. Trends in Cognitive Sciences, 24(12), 1028-1040. https://doi.org/10.1016/j.tics.2020.09.004
5. Johansson, F. D., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. In Proceedings of the 33rd International Conference on Machine Learning (Vol. 48, pp. 3020-3029). PMLR. https://proceedings.mlr.press/v48/johansson16.html
6. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwińska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521-3526. https://doi.org/10.1073/pnas.1611835114

7. Krawczyk, B., Minku, L. L., Gama, J., Stefanowski, J., & Woźniak, M. (2017). Ensemble learning for data stream analysis: A survey. Information Fusion, 37, 132-156. https://doi.org/10.1016/j.inffus.2017.02.004
8. Lopez-Paz, D., & Ranzato, M. (2017). Gradient episodic memory for continual learning. In Advances in Neural Information Processing Systems (Vol. 30). Curran Associates, Inc. https://arxiv.org/abs/1706.08840
9. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering, 31(12), 2346-2363. https://doi.org/10.1109/TKDE.2018.2876857
10. Montiel, J., Read, J., Bifet, A., & Abdessalem, T. (2018). Scikit-multiflow: A multi-output streaming framework. Journal of Machine Learning Research, 19(72), 1-5. https://www.jmlr.org/papers/v19/18-251.html
11. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. Neural Networks, 113, 54-71. https://doi.org/10.1016/j.neunet.2019.01.012
12. Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books.
13. Rabanser, S., Günnemann, S., & Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. In Advances in Neural Information Processing Systems (Vol. 32). Curran Associates, Inc. https://arxiv.org/abs/1810.11953
14. Ramamurthy, S. R., & Bhatnagar, A. (2007). Tracking recurrent concept drift in streaming data using ensemble classifiers. In Proceedings of the Sixth International Conference on Machine Learning and Applications (pp. 404-409). IEEE. https://doi.org/10.1109/ICMLA.2007.80
15. Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., & Hadsell, R. (2016). Progressive neural networks. arXiv. https://doi.org/10.48550/arXiv.1606.04671
16. Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 3076-3085). PMLR. https://proceedings.mlr.press/v70/shalit17a.html
17. Sethi, T. S., & Kantardzic, M. (2017). On the reliable detection of concept drift from streaming unlabeled data. Expert Systems with Applications, 82, 77-99. https://doi.org/10.1016/j.eswa.2017.04.008
18. Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L., & Petitjean, F. (2016). Characterizing concept drift. Data Mining and Knowledge Discovery, 30, 964-994. https://doi.org/10.1007/s10618-015-0448-4
19. Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. Machine Learning, 23, 69-101. https://doi.org/10.1007/BF00116900
20. Žliobaitė, I., Bifet, A., Pfahringer, B., & Holmes, G. (2014). Active learning with drifting streaming data. IEEE Transactions on Neural Networks and Learning Systems, 25(1), 27-39. https://doi.org/10.1109/TNNLS.2012.2236570