Interdisciplinary Sciences: Computational Life Sciences (2025) 17:390–408
https://doi.org/10.1007/s12539-024-00675-2

ORIGINAL RESEARCH ARTICLE

CEL: A Continual Learning Model for Disease Outbreak Prediction by Leveraging Domain Adaptation via Elastic Weight Consolidation

Saba Aslam[1,2] · **Abdur Rasool**[1,2] · **Xiaoli Li**[1,3,4] · **Hongyan Wu**[1]

Received: 2 January 2024 / Revised: 6 November 2024 / Accepted: 7 November 2024 / Published online: 28 February 2025

Abstract

Continual learning is the ability of a model to learn over time without forgetting previous knowledge. Therefore, adapting new data in dynamic fields like disease outbreak prediction is paramount. Deep neural networks are prone to error due to catastrophic forgetting. This study introduces a novel CEL model for Continual Learning by leveraging domain adaptation via Elastic weight consolidation (EWC). This model aims to mitigate the catastrophic forgetting phenomenon in a domain incremental setting. The Fisher information matrix (FIM) is constructed with EWC to develop a regularization term that penalizes changes to essential parameters. We conducted experiments on three distinct diseases, influenza, mpox, and measles, with customized metrics. The high *R*-squared values during evaluation and reevaluation outperform the other state-of-the-art models in several contexts. The results indicate that CEL adapts well to incremental data. CEL's robustness emphasizes its minimal 65% forgetting rate and 18% higher memory stability compared to existing benchmark studies. This study highlights CEL's versatility in disease outbreak prediction by addressing evolving data with temporal patterns. It offers a valuable model for proactive disease control with accurate and timely predictions.

✉ Hongyan Wu
hy.wu@siat.ac.cn

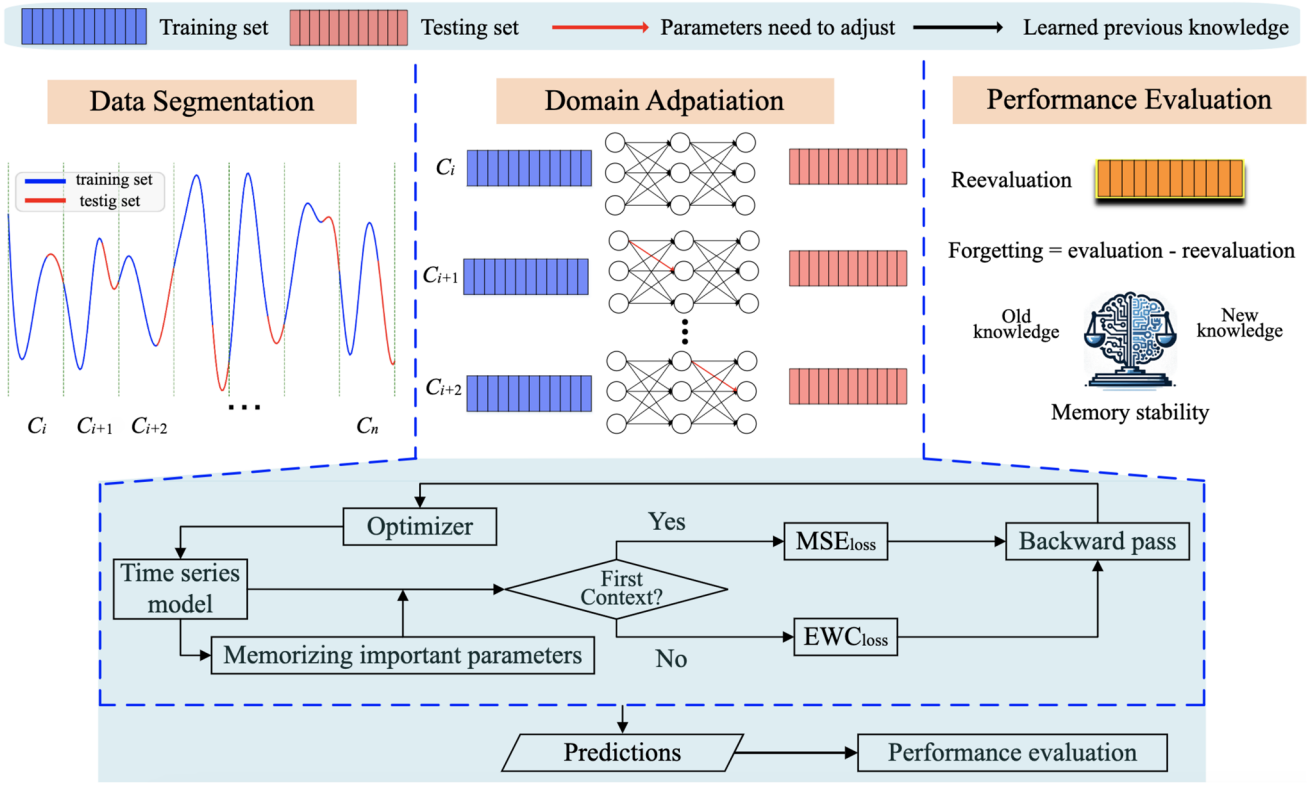
1 Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

2 Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Beijing 100049, China

3 Centre for Cognitive and Brain Sciences, University of Macau, Taipa, Macau SAR, China

4 Faculty of Science and Technology, University of Macau, Taipa, Macau SAR, China

Graphical Abstract



Keywords Continual learning · Domain-incremental learning · Domain adaptation · Disease outbreak prediction · Elastic weight consolidation

Abbreviations

CL	Continual Learning
CEL	Continual Learning via Elastic weight Consolidation
EWC	Elastic Weight Consolidation
FIM	Fisher Information Matrix
CF	Catastrophic Forgetting
IL	Incremental Learning
LSTM	Long Short-Term Memory
MSE	Mean Squared Error
Mpox	Monkey Pox
GEM	Gradient Episodic Memory
XdG	Context-dependent Gating
EMC	Episodic Memory systems

1 Introduction

Predicting disease outbreaks is crucial for timely healthcare response, resource allocation, and containment efforts. Early forecasts inform public awareness, guide research, and influence policy, benefiting public health and the economy. They also foster international collaboration in global health crises [1]. Deep learning algorithms have demonstrated their efficacy in predicting disease outbreaks [2]. Their hierarchical structure allows them to learn complex features from raw data, such as time-series data from disease surveillance or textual data from news reports [3]. However, these static deep learning models necessitate complete retraining when incorporating new data, leading to increased time, cost, and resource consumption. It highlights a deficiency in deep learning static models [4] because these models are often designed for specific tasks or static datasets and face challenges when it comes to evolving data like disease outbreak prediction.

In real time, disease outbreak data comes sequentially and should be incremented in the model based on daily, weekly, monthly, or yearly. In contrast, existing prediction models work statically, which means they fail for the purpose they are generated for [4]. Unlike artificial neural networks, entities like humans and animals can learn new things incrementally without forgetting previously learned skills. It emphasizes the need to improve artificial systems to mimic natural ones in this aspect [5].

In this scenario, continual learning (CL) has emerged as an effective solution. A continual learning system is an adaptive model that learns from a continuous stream of information over time without predefined task limits and forgetting or interference when accommodating new information [6]. Its primary goal is to develop models that allow artificial systems to learn continuously over time without forgetting previous knowledge [4, 7]. The difference between static and continual learning is illustrated in Fig. 1. The model (M_1) is trained on a fixed dataset (T_1) to give predictions (P_1) in static learning. When new data (T_2, T_3) is incremented, a new model (M_2, M_3) is needed to be retrained. In contrast, the CL model ($M_{1,1}$) can update its parameters in response to new data streams without retraining from scratch. For continual updates in time series prediction, traditional methods are often employed, which update models periodically with new data but do not pay much attention to previously learned knowledge [8]. In contrast, continual learning approaches [9] allow for the adaptation of model parameters in response to evolving data distributions.

Interest in CL is rising due to its vast potential applications with deep neural networks. These applications range from medical diagnosis [10, 11] (where new diseases or symptoms might be discovered) to autonomous driving [12] (where the driving environment and rules might change) and real-time video surveillance [13] (a high-dimensional application). Pre-trained language models BERT [14], ALBERT [15], RoBERTa [16], and GPT [17] have been successfully employed in continual learning of various natural language problems using different continual learning approaches [18, 19]. The mention of these applications emphasizes the practical importance and potential impact of advancements in CL. Across these diverse applications, the main challenge of catastrophic forgetting (CF) remains crucial. CF occurs when neural networks forget previously learned information upon learning new data. More concretely, as these networks learn new tasks or adapt to fresh data distributions, they often lose the ability to execute tasks they were originally trained on. This situation poses a significant challenge, mainly when the expectation is for models to seamlessly assimilate new data without compromising their expertise on prior tasks [7].

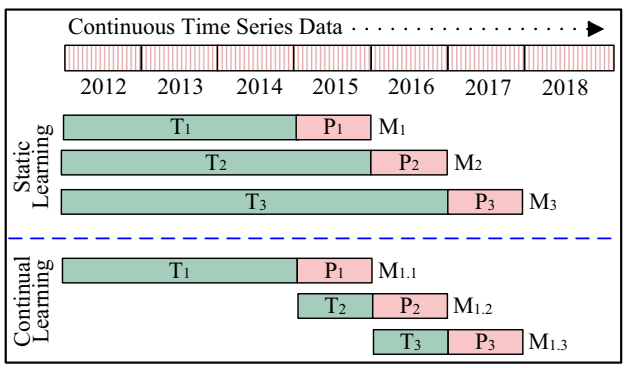


Fig. 1 Static learning vs continual learning, T represents training data, P predictions and M model

Numerous studies [20–23] have been developed to solve the CF problem. A fundamental understanding of CL's various types and scenarios is vital to devising robust solutions for combatting CF. It has three types [9]: (i) task-incremental learning (Task-IL), which is the sequential and incremental adaptation of an algorithm to a series of discrete tasks [24, 25]. (ii) class-incremental learning (Class-IL), where an algorithm is faced with the evolving challenge of differentiating an expanding array of objects or classes [23, 26, 27], and (iii) domain-incremental learning (Domain-IL) that describes a situation where the problem structure remains constant, but the context or input distribution shifts, leading to changes in domains [28]. However, most existing approaches are typically designed for Task-IL or Class-IL within the context of classification problems [9]. Meanwhile, regression tasks are essential in disease outbreak prediction, as predicting the number of patients is essential for vaccine production and healthcare resource management, such as bed allocation. Moreover, there is a notable absence of research in Domain-IL [9] explicitly targeting disease time-series prediction within the CL framework. It represents a gap in current research indicating potential avenues for further investigations.

In this study, we developed a novel model, CEL, for Continual Learning by leveraging domain adaptation via Elastic weight consolidation. It combats catastrophic forgetting in domain-incremental learning settings for disease outbreak prediction. The Fisher information matrix (FIM) is applied in EWC to construct a regularization term that penalizes changes to important parameters for previous knowledge. The process starts with data segmentation for contextual learning, followed by domain adaptation, where a neural network model incorporates with EWC and retains earlier knowledge while integrating new contexts. Finally, performance evaluation measures knowledge retention versus new learning. It enables CEL to maintain historical insights while staying updated with emerging data, optimizing its predictive accuracy for disease trends. The following significant contribution underscores the necessity of implementing our proposed CEL model in continual learning for disease outbreak prediction.

1) To the best of our knowledge, this research is pioneering in its attempt to develop a lightweight continual learning model for disease outbreak prediction by leveraging domain adaptation via EWC.
2) A novel data segmentation strategy is presented that partitions time-series data into discrete contexts, each

encapsulating a distinct data distribution to facilitate domain-incremental learning.

3) The results consistently showed the CEL's effective performance in maintaining high *R*-squared values, forgetting, and memory stability metrics outperforming the other state-of-the-art models.
4) This is the first study that focuses on regression problems within continual learning and introduces the adapted forgetting measures.

The rest of the paper is organized as follows: Sect. 2 reviews existing approaches, Sect. 3 describes the methodology, Sect. 4 outlines the experiments, Sect. 5 discusses the results, and Sect. 6 concludes the paper.

2 Literature Review

2.1 Deep Learning for Disease Outbreak Prediction

Deep neural networks have gained popularity among researchers in recent years for disease outbreak prediction. This is primarily due to their ability to extract meaningful features from input data [29]. Hybrid modeling approaches [30] employ specialized optimization techniques during model tuning and have demonstrated effectiveness in influenza prediction [31, 32]. Kara [31] employed genetic algorithms to optimize LSTM model hyperparameters, enhancing multi-step influenza prediction accuracy. On the other hand, Yang et al. [32] introduced a machine learning framework based on comprehensive learning particle swarm autoregression chains, which extracts patterns in various orders, resulting in improved performance compared to previous methodologies. Facebook has developed a Prophet model for time series prediction [33]. Prophet is designed for forecasting time series data based on an additive model where non-linear trends fit yearly, weekly, and daily seasonality and holiday effects. However, one of the main challenges with FB-Prophet is scalability when analyzing large datasets [34]. Twitter data is widely utilized in sales forecasting models, with numerous studies analyzing tweets and user engagement to predict sales trends on the platform [35, 36].

Weather forecasting models, essential in time series prediction, leverage advanced algorithms [37] and a mix of statistical and machine learning techniques [38, 39], demonstrating the field's sophistication and adaptability. Moreover, these deep learning models also suffer from catastrophic

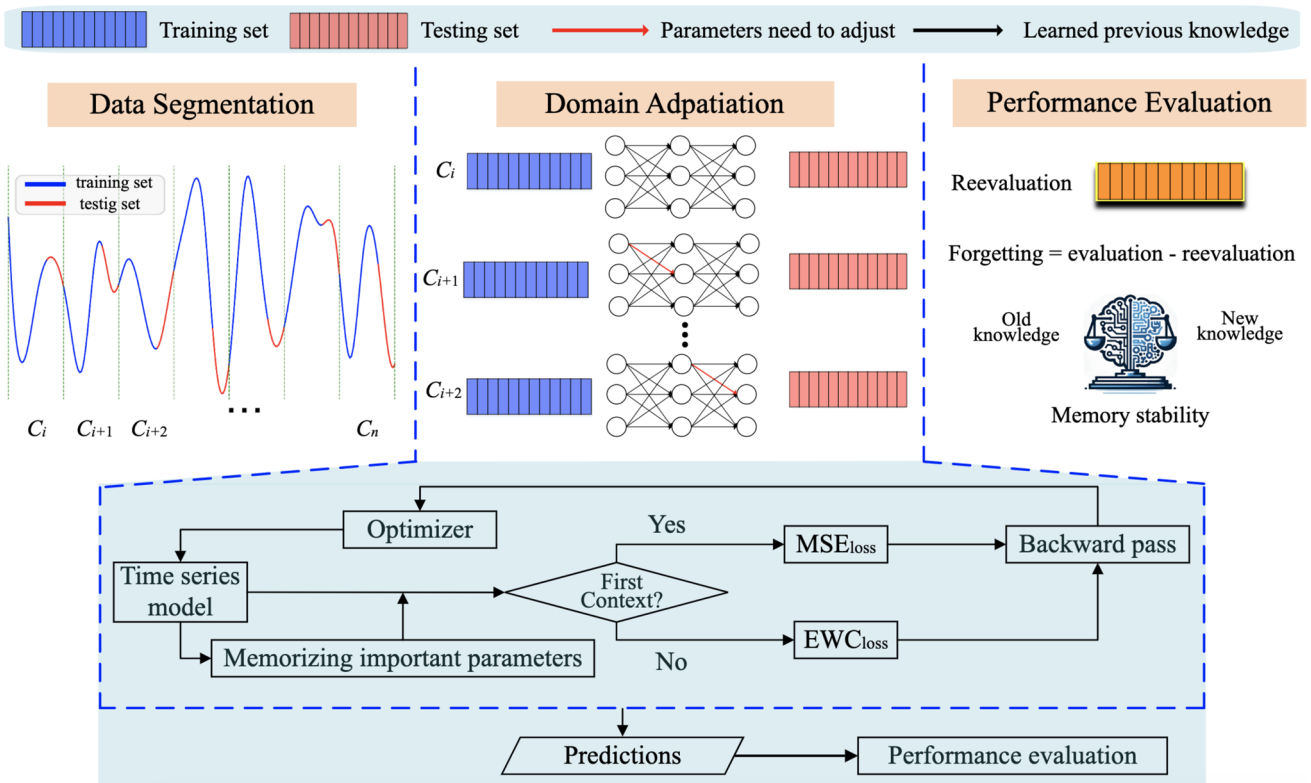


Fig. 2 The proposed CEL model takes disease time series data and predicts outbreak, where first data segmentation strategy segments the whole data into different contexts, then domain adaptation is presented where the model adapts to new data while retaining important previous knowledge and performance evaluation presents the customized metrics to assess the model's performance

forgetting when facing new and incremental data under continual learning [40].

2.2 Continual Learning

Approaches for three scenarios of continual learning, Task-IL, Class-IL, and Domain-IL, fall under three categories: replay-based, regularization-based, and parameter isolation. Replay-based approaches maintain a limited exemplar memory to rehearse prior tasks while training new ones [20, 41–43]. Generative models can replace this memory to simulate past distributions [21, 44]. Techniques exist to minimize interference between tasks [22, 24]. Regularization-based approaches add terms to the training loss to restrict weight changes. Knowledge distillation maintains previous learning while adapting to new tasks [23, 41, 45]. Another method involves calculating the importance of each parameter and updating them accordingly [40, 46]. Parameter isolation mainly operates without model size constraints. It isolates important parameters from prior tasks, enabling new parameters for new tasks [47–49]. Zero-forgetting is achieved through learning masks or paths for each parameter or layer [50, 51]. However, most existing methods in CL are developed for Task-IL or Class-IL, focusing on classification challenges, while Domain-IL and regression tasks remain largely neglected.

3 Methodology

3.1 Problem Definition

A significant characteristic of continual learning is its sequential learning process. We opted for a domain-incremental learning approach where the task (prediction) remains consistent, but data distribution undergoes shifts. At each time, only a tiny amount of the input data is available, called context C. Mathematically, Domain-IL for time series regression tasks can be expressed as follows:

Consider the number of N contexts $C = \left(C_1, C_2, \cdots, C_N\right)$ arriving sequentially. A context $C_i (i = 1,2, \cdots, N)$ consists of k instances of labeled disease data $\mathcal{D}_i = \left\{\left(x_{i,m}, y_{i,m}\right)\right\}_{m=1}^{k}$ with window size 7, each time-series $x_{i,m} \in \mathcal{X}_i$ with an associated target $y_{i,m} \in \mathcal{Y}_i$, and m refers to the m^{th} timepoint in the context C_i. The target space $\mathcal{Y}_i$ corresponds to real-valued numbers (or vectors). The goal in each context is to learn a solver model M_i with trainable parameters θ_i such that $M_i : \mathcal{X}_i \rightarrow \mathcal{Y}_i$, with an estimated target $\hat{y}_{i,m} = M_i\left(x_{i,m}; \theta_i\right)$ on condition of without accessing data from previous contexts $C_j (j = 1, 2, \ldots, i-1)$.

3.2 Proposed Model

The proposed CEL model mitigates catastrophic forgetting in Domain-IL, focusing on time-series data for multiple diseases. For a clear understanding of CEL, it is structured into three distinct phases, as depicted in Fig. 2. It illustrates a continual learning approach applied to a time series dataset for disease outbreak prediction. Initially, to simulate the model's exposure to evolving data, we segmented the disease time-series data into various distinct contexts. Each context is further divided into a training set (shown in blue) and a testing set (shown in red). During the training phase, we start by feeding the time series model. The model memorizes the important parameters for each context by introducing a regularized loss, balancing the acquisition of new information from the current context with the retention of essential knowledge from previous ones. Performance is evaluated based on the learned knowledge from all the contexts with the different metrics. Overall, the proposed methodology is divided into the following subsections.

i. *Data Segmentations*: We introduced a segmentation strategy by dividing the entire dataset into discrete N contexts, each of which was split into both 80% training and 20% testing sets explained in Sect. 3.2.1.
ii. *Domain Adaptation*: Our CEL model uses EWC and FIM by overcoming catastrophic forgetting and facil-

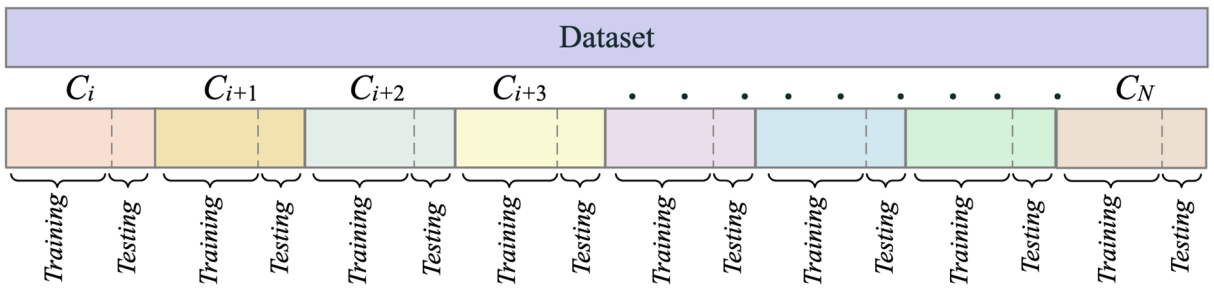


Fig. 3 Data segmentation strategy segments the time series data into different contexts

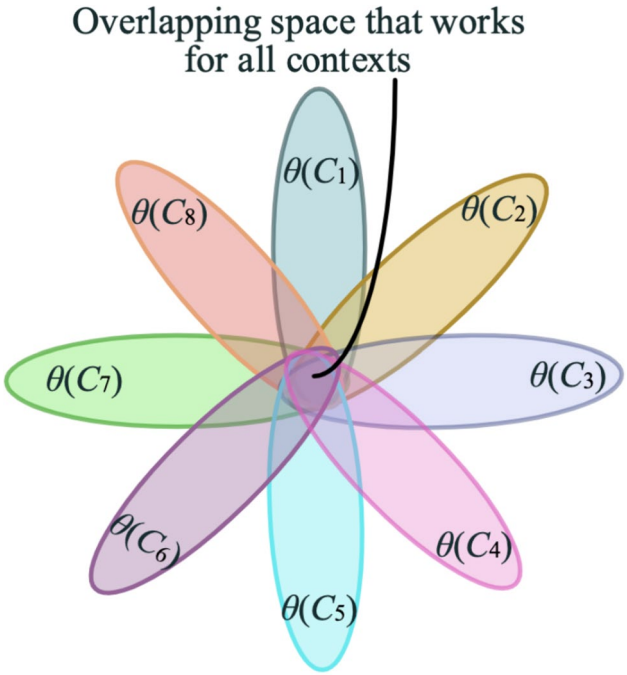


Fig. 4 An example of eight contexts where EWC ensures that old contexts are remembered while training on new ones

itating domain adaptation to new data described in Sect. 3.2.2.

iii. *Performance Evaluation*: We customized continual learning performance evaluation metrics [52] to assess the model's performance elaborated in Sect. 3.2.3.

The following sub-sections explain the proposed model step by step.

3.2.1 Data Segmentation

We adopted a segmentation strategy where we meticulously segmented our complete dataset into distinct N contexts to address the challenge posed by evolving data distributions. Each context C_i ($i = 1,2, \cdots, N$, where N is the total number of contexts in the dataset) representative of unique data distribution was further divided into training (80%) and testing (20%) sets, as illustrated in Fig. 3. This segmentation strategy has two following objectives:

- We simulate the model's exposure to evolving data distributions by introducing diverse contexts and subjecting the model to training and testing phases within each context. It robustly gauges and augments its adaptability.
- This bifurcation ensures that as the model processes new data, its knowledge retention is periodically validated against the testing set, promoting consistent learning while preventing overfitting.

Time series disease data inherently shows variations, both temporally and geographically. Such a strategy is imperative for deriving comprehensive and accurate insights into the trajectory of the disease. This segmentation strategy revolves around selecting the optimal number of contexts N. For this, we partitioned the complete dataset into $N(= 6,7, \ldots, 10)$ contexts, executed a computational model on these segmented contexts, computed the average R-squared values for each configuration, and finalized N value satisfying the following.

i. More contexts allowed our model to capture the nuances in the data distribution, resulting in improved predictive performance.
ii. Do not unnecessarily burden the model with an excessive number of contexts while still enabling effective learning.

3.2.2 Domain Adaptation

In this subsection, we mitigate CF phenomenon by leveraging domain adaptation to new data. The problem structure remains constant in domain-incremental learning, but the context or input distribution shifts. Therefore, we need to leverage domain adaptation to new data while preventing the catastrophic forgetting of the knowledge from the older context. Drawing inspiration from the theory of plasticity of post-synaptic dendritic spines in the brain [5], EWC introduces a paradigm that determines the importance of a network parameter to previous contexts and provides a solution to the continual learning challenge by consolidating context-specific synaptic strengths (network parameters) [40]. It then penalizes any changes made to this parameter based on its importance while learning new contexts. To explore how EWC addresses the challenges of approximating the posterior of θ in the context of deep learning models, this subsection paves the way for more effective learning across multiple contexts. The training of deep neural networks is fundamentally an optimization problem. The objective is to find an optimal set of parameters, denoted by θ, that minimizes the error in the training objective. Given the complexity of real-world contexts, there often exist multiple configurations of θ that can achieve near-optimal performance. In a CL scenario, where multiple contexts are learned sequentially, the challenge is to find a configuration of θ that performs well across all contexts. It requires the network to select parameters from the overlapping solution space of all individual contexts [53], as represented in Fig. 4.

Memorizing Important Parameters

To explore how EWC addresses the challenges of approximating the posterior of θ in the context of deep learning models, this subsection paves the way for more

effective learning across multiple contexts. The training of deep neural networks is fundamentally an optimization problem. The objective is to find an optimal set of parameters, denoted by θ, that minimizes the error in the training objective. Given the complexity of real-world contexts, there often exist multiple configurations of θ that can achieve near-optimal performance. In a CL scenario, where multiple contexts are learned sequentially, the challenge is to find a configuration of θ that performs well across all contexts. It requires the network to select parameters from the overlapping solution space of all individual contexts [53], as represented in Fig. 4.

To regularize the appropriate parameters, EWC determines the importance of each parameter to that context once a context C_i is learned. This importance is quantified using FIM. FIM [54] captures the learned probabilistic model's sensitivity to parameter changes. FIM (F_i) is typically calculated as the second derivative of the log-likelihood function with respect to the model parameters. In mathematical notation, it can be expressed as [53]

$$F_i(\theta) = \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log\rho(x \mid \theta)\right] \tag{1}$$

where $F_i(\theta)$ is the FIM regarding the model parameters θ, $\mathbb{E}$ presents the expectation, which is usually calculated over the entire dataset, and $\frac{\partial^2}{\partial\theta^2}\log\rho(x \mid \theta)$ is the second derivative of the log-likelihood function with respect to the model parameters θ.

Contextual Loss Formulation

After learning a context, the network parameters are fine-tuned for subsequent contexts. However, changes to parameters deemed important for previous contexts are penalized. The penalty is proportional to the squared difference between the new and old values of the parameter, weighted by its importance. The new objective function, which includes this penalty, is [40]

$$L(\theta) = L_{\text{new}}^{\text{original}}(\theta) + \sum_i \lambda/2\left[F_i\left(\theta_i - \theta_{i,\text{old}}^*\right)^2\right] \tag{2}$$

where $L_{\text{new}}^{\text{original}}(\theta)$ is the original loss function for a new context, λ is a hyperparameter that determines the strength of the regularization. F_i is the Fisher information for parameter θ_i, which quantifies its importance to the old context. $\theta_{i,\text{old}}^*$ is the value of the parameter θ_i after training on old context. The term $\sum_i \lambda/2\left[F_i\left(\theta_i - \theta_{i,\text{old}}^*\right)^2\right]$ acts as a regularization term. It penalizes changes to parameters that were important for the old context. The larger the Fisher information F_i for a parameter, the more it is penalized for deviating from its value after training in the old context.

In essence, while training on a context, the network tries to minimize the error and ensures that it doesn't forget the knowledge from the old context. This is achieved by adding a penalty to the loss function, discouraging significant changes to important parameters from the old context. This loss function is central to the EWC method and ensures the network can learn new contexts without forgetting the old ones.

Time Series Prediction

We chose long short-term memory (LSTM) [31] as a deep learning model for disease outbreak prediction because LSTM model is specifically designed to handle sequential data, making them particularly suitable for lightweight time series continual learning.

In LSTM cells, a crucial component termed the cell state c_{t-1} acts as the long-term memory of the unit. Another vital aspect of the LSTM unit is the hidden state h_{t-1}, which symbolizes the short-term memory. In the sigmoid activation function, the output effectively dictates the proportion of retained long-term memory. It fundamentally influences the retention capacity of long-term memory and is aptly termed the forget gate.

$$f_l = \sigma\left(W_{l,h}h_{t-1} + W_{l,x}x_t + b_l\right) \tag{3}$$

where f_l is the percentage of long-term to remember by forget gate.

Synthesis of potential long-term memory involves assimilating short-term memory and the current input. Concurrently, the network must determine the quantum of this potential memory to be incorporated into the long-term memory, which is referred to as the input gate.

$$p_i = \text{sigmoid}\left(W_{p,h}h_{t-1} + W_{p,x}X_t + b_p\right) \tag{4}$$

$$l_i = \tanh\left(W_{i,h}h_{t-1} + W_{i,x}X_t + b_i\right) \tag{5}$$

where p_i is the percentage of potential memory to remember by input gate and l_i is the potential long-term memory by the input gate.

Following the acquisition of a new long-term memory, the new long-term memory transforms the tanh activation function. The resultant value then dictates the proportion of short-term memory to be propagated forward, and this phase is labeled as the output gate.

$$c_t = c_{t-1}f_l + p_i l_i \tag{6}$$

$$p_o = \sigma\left(W_{o,h}h_{t-1} + W_{o,x}x_t + b_o\right) \tag{7}$$

$$h_t = p_o\tanh\left(c_t\right) \tag{8}$$

where c_t is the new long-term memory or termed as the new cell state, p_o is the percentage of the potential memory to remember by output gate and h_t is the new short-term memory which is the output.

LSTM networks use consistent weights and biases across stages to handle varying data sequence lengths. At each step, LSTMs learn weight parameters W and bias parameters b to reduce prediction discrepancies.

CEL Training

Initially, we segmented the disease time-series data into various distinct contexts (subSect. 3.2.1). Each of these contexts represents a unique environment or domain where the aim is to predict the outbreak of a particular disease. We feed the time series model with the first context during the training phase. This model generates predictions and the mean squared error (MSE), which is the original loss function for this first context. Mathematically, this can be represented as [55]:

$$l_{\text{MSE}} = \frac{1}{k}\sum_{m=1}^{k}\left(y_m - \acute{y}_m\right)^2 \tag{9}$$

where y_m and $\acute{y}_m$ are the actual and predicted values for the m^{th} instance and k is the total number of instances in C_i.

Once the initial training on the first context is completed, the model parameters are saved, and the FIM (subSect. 3.2.2, Memorizing Important Parameters) evaluates the importance of each model parameter. Essentially, FIM is calculated as the expected value of the square of the gradient of the loss function concerning each parameter. The formula for a fundamental FIM element is represented in Eq. (1). EWC (subSect. 3.2.2, EWC) calculates the penalty term. The FIM values guide the penalty term. Thus, for subsequent contexts, the loss function evolves into a regularized loss (L'), as follows, that incorporates both the original loss (MSE) and a penalty term P [40].

$$L' = l_{\text{MSE}} + P \tag{10}$$

$$P = \sum_i \lambda/2\left[F_i\left(\theta_i - \theta^*_{i,\text{old}}\right)^2\right] \tag{11}$$

where λ is a regularization parameter, and $\theta_{i,\text{ old}}$ represents the parameters from the old contexts.

This learning process is iterative and continues for each subsequent context. Model parameters are consistently updated based on their weighted importance from previous contexts, as indicated by the FIM values. By adhering to this approach, CEL adapts to new data while retaining the knowledge acquired in earlier contexts, thereby balancing adaptability and memory retention. All these steps are presented in Algorithm 1.

3.2.3 Performance Evaluation Metrics

We customized the following performance metrics [52] for the CEL evaluation. These metrics are theoretically varied and explained in our scenario.

R-squared during evaluation: During the evaluation phase (concurrent with the training phase of sequential data), each context C_i is evaluated on its respective test set after being trained on the cumulative training sets of all previous contexts $\sum_{j=1}^{i} C_j$. The R-squared (R^2) for each context is calculated as

$$R^2_{\text{EV},C_i} = 1 - \frac{\sum_{m=1}^{k}\left(y_m - \acute{y}_m\right)^2}{\sum_{m=1}^{k}\left(y_m - \bar{y}\right)^2} \tag{12}$$

where y_m and $\acute{y}_m$ are the actual and predicted values for the m^{th} instance, k is the total number of instances, and $\bar{y}$ denotes the mean of the observed values in the test set of C_i. R-squared takes any values between 0 and 1, and a higher R-squared indicates that the model explains more variability.

R-squared during reevaluation: After the model has been trained on all contexts $\sum_{i=1}^{N} C_i$ (where N is the total number of contexts in the dataset), its performance is reevaluated on the test set of each context. The R-squareduring reevaluation

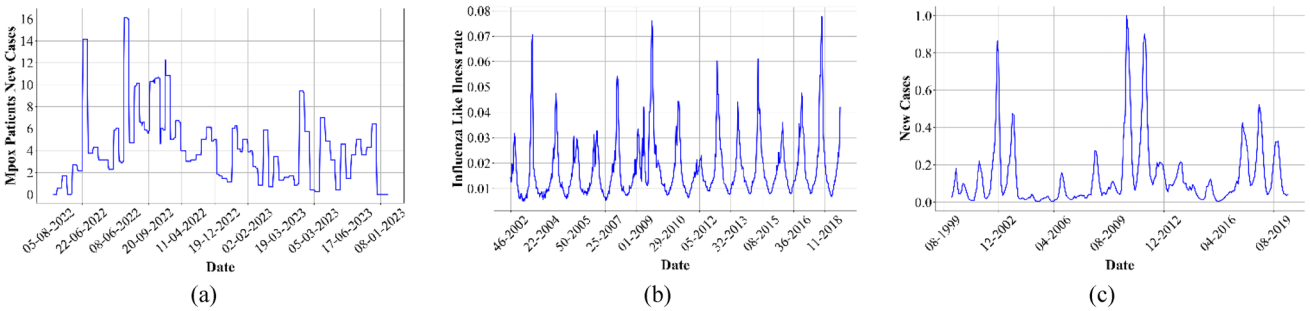


Fig. 5 Visualization of different disease data, **a** daily mpox cases in Africa, **b** weekly influenza-like illness rate in the US, **c** monthly measles cases in the EU

Table 1 Context-based data description of mpox data

Context	Start Date	End Date	Days	Mean	Std
1	08-05-2022	21-06-2022	45	2.68	4.23
2	22-06-2022	05-08-2022	45	4.20	2.45
3	06-08-2022	19-09-2022	45	8.02	4.24
4	20-09-2022	03-11-2022	45	7.40	2.77
5	04-11-2022	18-12-2022	45	4.14	1.28
6	19-12-2022	01-02-2023	45	3.43	1.84
7	02-02-2023	18-03-2023	45	2.44	1.79
8	19-03-2023	02-05-2023	45	3.03	3.28
9	03-05-2023	16-06-2023	45	3.43	2.27
10	17-06-2023	31-07-2023	45	3.48	2.21

Table 2 Context-based data description of influenza data

Context	Start Date	End Date	Weeks	Mean	Std
1	46-2002	25-2004	84	0.015	0.014
2	26-2004	May-2006	84	0.015	0.009
3	22-2005	36-2007	84	0.014	0.008
4	37-2007	16-2009	84	0.018	0.011
5	17-2009	48-2010	84	0.021	0.016
6	49-2010	28-2012	84	0.016	0.009
7	29-2012	Jul-2014	84	0.019	0.012
8	32-2013	39-2015	84	0.017	0.011
9	40-2015	19-2017	84	0.020	0.010
10	20-2017	50-2018	84	0.020	0.018

Table 3 Context-based data description of measles data

Context	Start Date	End Date	Months	Mean	Std
1	Aug-1999	Nov-2002	39	0.080	0.060
2	Dec-2002	Mar-2006	39	0.238	0.232
3	April-2006	Jul-2009	39	0.021	0.010
4	Aug-2009	Nov-2012	39	0.043	0.042
5	Dec-2012	Mar-2016	39	0.091	0.070
6	April-2016	Jul-2019	39	0.395	0.320

for each context is measured similarly to the *R*-squared during an evaluation.

$$R^2_{\mathrm{RE},C_i} = 1 - \frac{\sum_{m=1}^{k} \left(y_m - \acute{y}_m\right)^2}{\sum_{m=1}^{k} \left(y_m - \bar{y}\right)^2} \tag{13}$$

where R^2_{RE,C_i} is the *R*-squared value of C_i during reevaluation. This helps to understand if there has been catastrophic forgetting.

R-squared Forgetting: It measures how much performance has dropped (if at all) for each context after the model has been trained on all contexts $\sum_{i=1}^{N} C_i$. *R*-squared forgetting for each context C_i is calculated as

$$F_{R^2}(C_i) = R^2_{\mathrm{EV},C_i} - R^2_{\mathrm{RE},C_i} \tag{14}$$

where R^2_{EV,C_i} (Eq. (12)) is *R*-squared of C_i during evaluation and R^2_{RE,C_i}(Eq. (13)) is the *R*-squared during reevaluation.

It ranges from [−1,1], where 1 represents the higher forgetting, showing the model has forgotten everything about the C_i after learning a new context, 0 implies no forgetting, and −1 suggests improvement in the context C_i after learning new contexts.

Memory Stability: Another forgetting measure is memory stability. It's the complement of the average forgetting. It indicates how stable the learning is across all contexts $\sum_{i=1}^{N} C_i$. Memory stability of *R*-squared can be defined as

$$S_{R^2} = 1 - \frac{1}{N} \sum_{i=1}^{N} F_{R^2}(C_i) \tag{15}$$

The range of the values is [0,1], where 1 show that the model is highly stable in remembering previous knowledge, and 0 indicates the opposite, where the model is unstable in retaining past knowledge.

Table 4 Information on the experimental process values

Category	Name (value or version)
Hyperparameter optimal values	Learning rate (0.01), Hidden dimension (32). Batch size (32), Lambda (1000), Epochs (200)
Experimental environment	Intel(R)Core (TM) i7-6700 CPU @ 3.40GHZ3.41 GHz and 16 GB RAM
Software & Libraries	PyTorch (1.9), Numpy (1.24.3), Pandas (2.0.3), Scikit-learn (1.3.0), Matplotlib (3.7.2), Seaborn (0.12.2), NumPy (1.24.3)

Algorithm 1 The proposed CEL model

Input: Context (C), Number of contexts *(N)*, Train data (T_r), Test data (T_s), ewc_lambda (λ), Fisher information matrix (FIM).
LSTM model initialization
Set loss == MSE
$C_N \leftarrow$ train_test_split_contexts(data, N)
for C_i in C_N **do**
  T_r, $T_s \leftarrow$ load_data(C)
  for C_i (T_r, T_s) **do** //evaluation
    for each feature and label in T_r **do**
      predictions $\leftarrow$ model(inputs)
if ($C_i == 0$) **then**
  loss $\leftarrow$ criterion(predictions, labels)
else
  loss $\leftarrow$ EWC_loss(predictions, labels, λ, LSTM, FIM)
end if, end for
FIM $\leftarrow$ compute_FIM(T_r)
for C_i (T_r, T_s) **do** //reevaluation
  for each feature and label in T_s **do**
  predictions $\leftarrow$ model(inputs)
  end for
  Compute R-squared
end for
for C_i in C_N **do**
  Compute forgetting
end for
Calculate memory stability
Output: Trained CEL model with predictions

4 Experiments

4.1 Data Acquisition and Preprocessing

Monkey Pox (mpox) Dataset

Mpox disease outbreak data was acquired from the official repository of "Our World in Data" [56]. The new daily confirmed cases from the geographical region of Africa have a timeframe from 08 May 2022 to 31 July 2023, which is approximately 15 months. Post this initial preprocessing, we derived a subset of the data termed smooth cases from refining our dataset further and making it amenable for CL model. Figure 5a provides a time-series visual representation of the number of mpox disease patients. According to the segmentation strategy, a description of the context-based data is presented in Table 1.

Influenza Dataset

We sourced data on influenza-like illness (ILI) rates in the United States spanning from week 30 of 2003 to week 51 of 2019 from the Centers for Disease Control and Prevention's National Center for Immunization and Respiratory Diseases via their "FluView Interactive" platform [57]. This dataset comprises approximately 832 weekly data points, representing a comprehensive temporal record over nearly 16 years. Figure 5b is a time-series visual representation of these ILI rates, providing insights into overarching trends and seasonality. Table 2 presents the context-wise data description according to the segmentation strategy.

Measles Dataset

The dataset pertaining to measles incidence was meticulously procured from the European Centre for Disease Prevention and Control's Atlas platform [58]. This data encompasses monthly observations for the European Union (EU) region from January 2001 to December 2019, depicted in Fig. 5c. The whole data was divided into different contexts according to the segmentation strategy; the descriptions of these contexts are given in Table 3. Prior to its integration into our continual learning model, this dataset underwent a normalization process using min–max normalization to ensure uniformity and compatibility [31].

$$\tilde{x} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{16}$$

where $\tilde{x}$ denotes the normalized value, and x represents the actual value.

4.2 Experimental Implementation

Data for mpox, influenza, and measles were collected and preprocessed as detailed in Sect. 4.1, which includes data sources and preprocessing methods. The datasets were then segmented into N contexts. Each context was divided into 80% training and 20% testing sets. For the mpox and influenza, $N = 10$ and Measles $N = 6$ contexts were selected by segmentation strategy. All models (GEM [42], XdG [47], EMC [41], LSTM [31], Transformer [59], and CEL) were sequentially trained across each context for disease predictions by testing on its corresponding test set utilizing Eq. (12). Reevaluation was performed following the completion of the training across all contexts based on Eq. (13). Additionally, forgetting (F_{R^2}) using Eq. (14) and memory stability based on Eq. (15) were also calculated.

We employed a grid search approach to explore the hyperparameter space based on its impact on the model's performance metric, R-squared. The model was trained and evaluated across multiple contexts for each combination of hyperparameters. The performance was then averaged over all contexts to determine the most effective set of hyperparameters, and the optimal values were found, which are presented in Table 4. It also included the experimental environment, software, and libraries.

4.3 Baseline Models

We compared our proposed CEL model with the following studies. We selected these studies because of their distinct CL strategies.

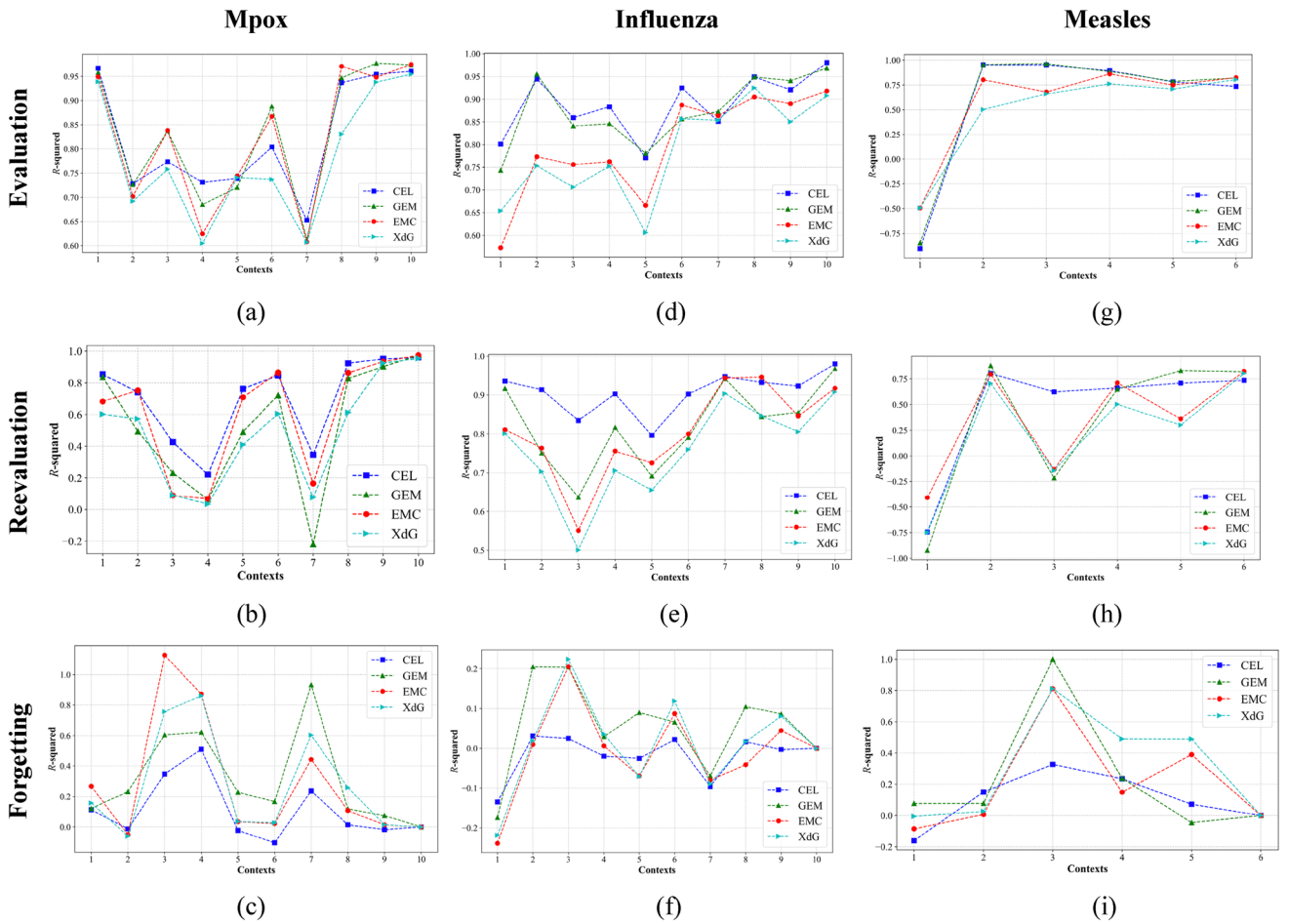


Fig. 6 *R*-squared trends during evaluation, reevaluation, and forgetting: **a** evaluation of mpox, **b** reevaluation of mpox, **c** forgetting of mpox, **d** evaluation of influenza, **e** reevaluation of influenza, **f** forgetting of influenza, **g** evaluation of measles, **h** reevaluation of measles, **i** forgetting of measles

- *GEM:* Operating on the Gradient Episodic Memory (GEM) principle, this model combats forgetting by retaining a subset of data from earlier contexts. Its model uses MLP layers and ensures that model updates don't adversely affect performance on previously stored data by suitably constraining gradient updates [42].
- *XdG:* This model uses a gating mechanism within its RNN structure using the Context-dependent Gating (XdG) strategy. It modulates RNN activations based on the context at hand. It tailors its learning to the current context without compromising knowledge from previous contexts by "gating" specific activations [47].
- *EMC:* EMC uses episodic memory systems and prediction-error-driven memory consolidation. Its model architecture uses adaptive architecture and utilizes the LSTM layers to mitigate catastrophic forgetting [41].

4.4 Traditional Models

We have selected the following traditional models to compare the CEL significance against them.

- *LSTM:* It handles sequential data and time series analysis efficiently to understand the temporal dynamics. Its unique architecture of memory cells and gates regulates the flow of information to remember it and forget irrelevant data, thereby mitigating the vanishing gradient problem [31].
- *Transformer*: The unique self-attention mechanism allows it to process sequential data in parallel, offering a significant advantage in capturing complex temporal dependencies compared to traditional recurrent neural networks. We adapted the original transformer architecture [59] for our purpose of time series prediction.

Table 5 Comparisons of memory stability of different models

Disease	GEM [42]	EMC [41]	XdG [47]	CEL
Mpox	0.7062	0.6899	0.6188	0.86213
Influenza	0.9088	0.8972	0.8061	0.96265
Measles	0.7597	0.7304	0.6621	0.84259

All baseline and traditional models were evaluated against CEL under consistent experimental settings.

5 Results and Discussion

In disease outbreak prediction, the ability of a model to adapt to new data without forgetting previously learned patterns is paramount. This work evaluated the proposed CEL model and compared it with GEM [42], XdG [47], EMC [41], LSTM [31], and Transformer [59].

5.1 Outbreak Predictions and Comparisons

We selected $N = 10$ contexts for mpox and influenza and $N = 6$ contexts for Measles for the segmentation strategy. We sequentially trained all models (GEM [42], XdG [47], EMC [41], and CEL) across each context for these disease predictions. After training on each training set of contexts, each model was assessed on its corresponding test set utilizing Eq. (12). Reevaluation was performed following the completion of the training across all contexts based on Eq. (13). Additionally, forgetting (F_{R^2}) using Eq. (14) and memory stability based on Eq. (15) were also calculated. A comparative analysis of the performance metrics of these models is presented in subsequent sections.

5.1.1 Mpox Prediction

We computed the *R*-squared values during the evaluation phase, depicted in Fig. 6a. It indicates the models' aptitude in learning the training data. A noticeable enhancement in performance was observed across all models during this phase; CEL consistently delivered outstanding *R*-squared scores across multiple contexts. Other models also demonstrated commendable *R*-squared values that CEL occasionally outperformed. After reevaluation, the *R*-squared values were again calculated and presented in Fig. 6b. It was observed that, apart from CEL, all other models demonstrated diminished performance. For instance, during task 7, CEL achieved the highest reevaluation score of 0.3454, whereas GEM recorded the lowest score of − 0.2192. This suggests that while these models were proficient during the training phase, they appeared to forget information from prior contexts when subjected to training in subsequent contexts. Further analysis was conducted to assess the models' retention capabilities illustrated in Fig. 6c, wherein CEL was found to exhibit the least amount of forgetting, as a comparative analysis revealed that CEL outperformed EMC by 69% at task 3 underscoring its superior retention abilities.

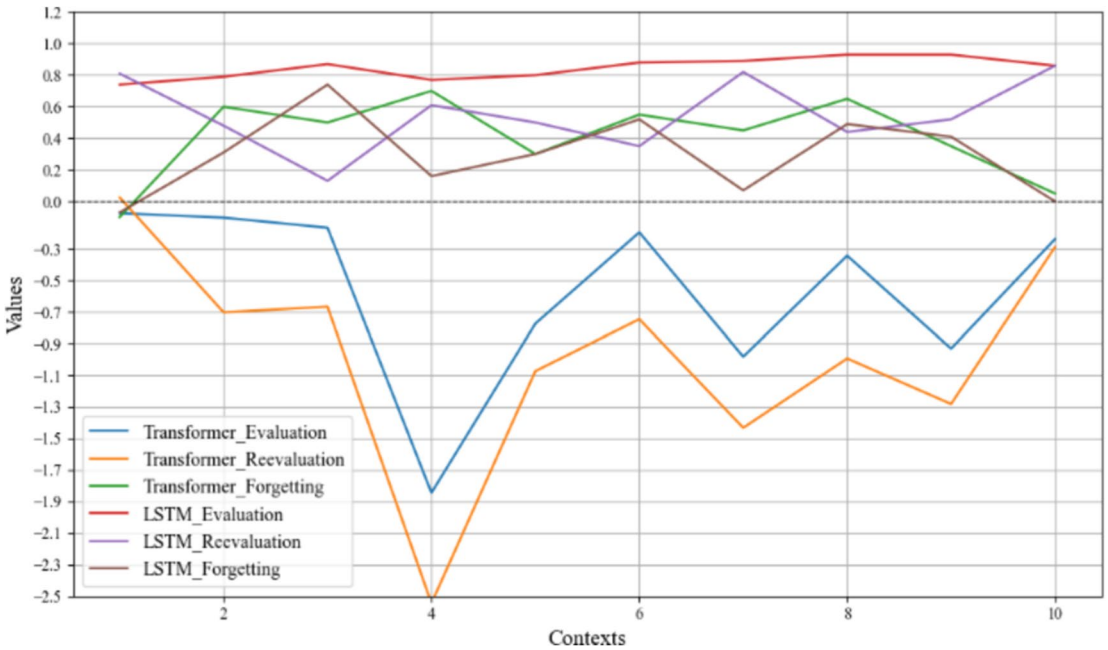


Fig. 7 *R*-squared trends during evaluation, reevaluation, and forgetting of LSTM and Transformer

Table 6 Set of parameters utilized in domain adaptation

Parameter Name	Number of Parameters	Transformation Type
lstm.weight_ih_l0	1536	Input to Hidden
lstm.weight_hh_l0	4096	Hidden to Hidden
lstm.bias_ih_l0	128	Input to Hidden
lstm.bias_hh_l0	128	Hidden to Hidden
linear.weight	32	Fully Connected
linear.bias	1	Fully Connected

5.1.2 Influenza Prediction

CEL displayed elevated *R*-squared values across every context during evaluation, suggesting its adeptness in adapting to the training influenza data, as demonstrated in Fig. 6d. Interestingly EMC manifested progressive adaptability, with their efficacy becoming increasingly apparent in later contexts. When these models were reevaluated, CEL maintained its proficiency, underscoring its robust generalization capabilities, presented in Fig. 6e. For example, during task 3, XdG obtained the lowest score at 0.5005, whereas CEL exhibited a notable peak at 0.8344. It suggests that despite their strong training performance during evaluation, other models exhibited a drop during reevaluation proving their potential overfitting during training. The forgetting analysis Fig. 6f revealed CEL's prowess in retaining knowledge from previous contexts, with minimal deterioration in performance.

In contrast, all other models displayed more pronounced forgetting in contexts, suggesting high forgetting in maintaining the previous knowledge. For example, in task 3, CEL exhibited an 88% higher proficiency in preserving information from prior contexts than the average performance of all other models.

5.1.3 Measles Prediction

CEL demonstrated a diverse performance, with notably high *R*-squared values during the training phase when models were trained on the Measles dataset, as illustrated in Fig. 6g. Other models also showed consistent and commendable performance across all contexts. These models were reevaluated to evaluate in terms of retention and adaptability. Results presented in Fig. 6h show that CEL's performance indicated its robustness in generalizing. The robustness of this performance is evident in task 3, where CEL (0.6245) outperformed all other models by 481%, showing a substantially superior performance compared to the average score of − 0.1636, which indicates an inconsistent outcome. These results exhibited a decline in showing some degree

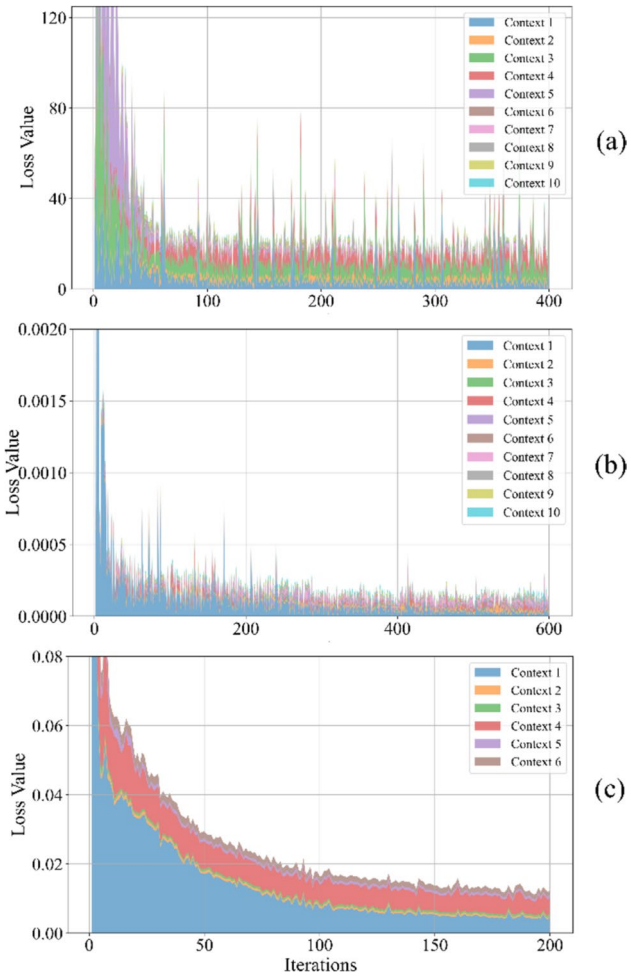


Fig. 8 Regularized loss by proposed CEL model over different contexts of mpox (**a**), influenza (**b**), and measles datasets (**c**)

of forgetting in all other models except CEL, suggesting potential challenges in retaining knowledge. The forgetting analysis in Fig. 6i further highlights the performance of the models in terms of forgetting. It shows that CEL was relatively stable throughout the context, while other models showed most forgetting. For example, in task 5, CEL exhibited proficiency (0.0716) in memorizing previous contexts, whereas XdG displayed a context-forgetting rate of 0.4892.

Across all three diseases, the proposed model, CEL, consistently demonstrated robust and stable performance, both during the evaluation and reevaluation phases. While generally performing well, GEM faced challenges in specific contexts, indicating potential areas for optimization. EMC and XdG, though showing promise in several contexts, had their struggles, especially in the initial contexts for each disease. In analyzing the average trend of forgetting across all tasks within three disease datasets for four models, it is evident

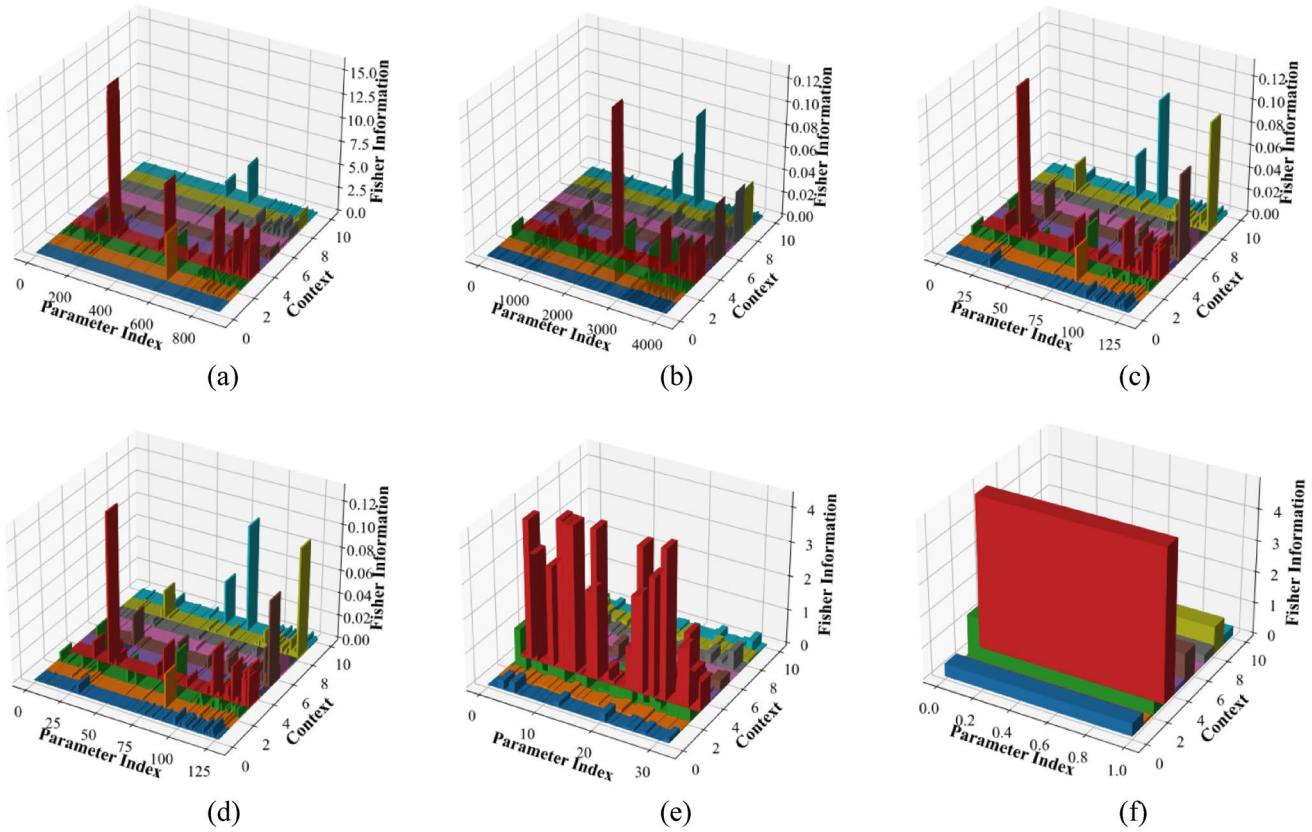


Fig. 9 Fisher information changes for different parameters during the training of mpox data by CEL, **a** lstm.weight_ih_10, **b** lstm.weight_hh_10, **c** lstm.bias_ih_10, **d** lstm.bias_hh_10, **e** linear.weight, and **f** linear.bias

that the proposed CEL model demonstrates the 65% lowest average forgetting, comparatively to other models.

Regarding memory stability, a crucial metric in CL, CEL emerged with superior performance, as in Table 5. Our findings indicate that in the mpox data, CEL demonstrated a 39% superior resilience in retaining knowledge compared to XdG. Furthermore, it proved to be 17% and 19% more effective in ensuring memory stability than the average performance of other models in the influenza and Measles datasets, respectively. The analysis of overall memory stability becomes evident through a comparison of the average performances of each model across all disease datasets, with the following hierarchical order: CEL > GEM > EMC > XdG.

5.2 Traditional Model Results

In experiments, we also predict influenza outbreak with traditional LSTM [31] and Transformer [59] models to reference how traditional models work in a continual learning setting. We chose influenza prediction considering that influenza has more data and is one of the typical epidemic diseases.

In Fig. 7, LSTM performed well during evaluation, with a mean R-squared value of 0.846 and a standard deviation of 0.064, indicating sophisticated adaptation to the training influenza data. However, its performance declined significantly during reevaluation, suggesting potential overfitting during training. Additionally, LSTM displayed pronounced forgetting across contexts, indicating difficulty retaining previous knowledge with the lowest 0.71 memory stability.

In a continual learning setting, models must update their parameters with typically limited new data. This scenario is less favorable for transformers, which are optimized for training with larger datasets. Consequently, our results (Fig. 7) indicate that Transformer evaluation underperformed compared to LSTM evaluation in all 10 contexts, with an average performance gap of – 1.541, and underperformed in reevaluation in 9 contexts, with an average performance gap of – 1.623.

5.3 FIM Impact

The FIM assumes a pivotal role in CEL's disease outbreak prediction model, offering insights into how distinct model components contribute to predicting disease outbreaks. We

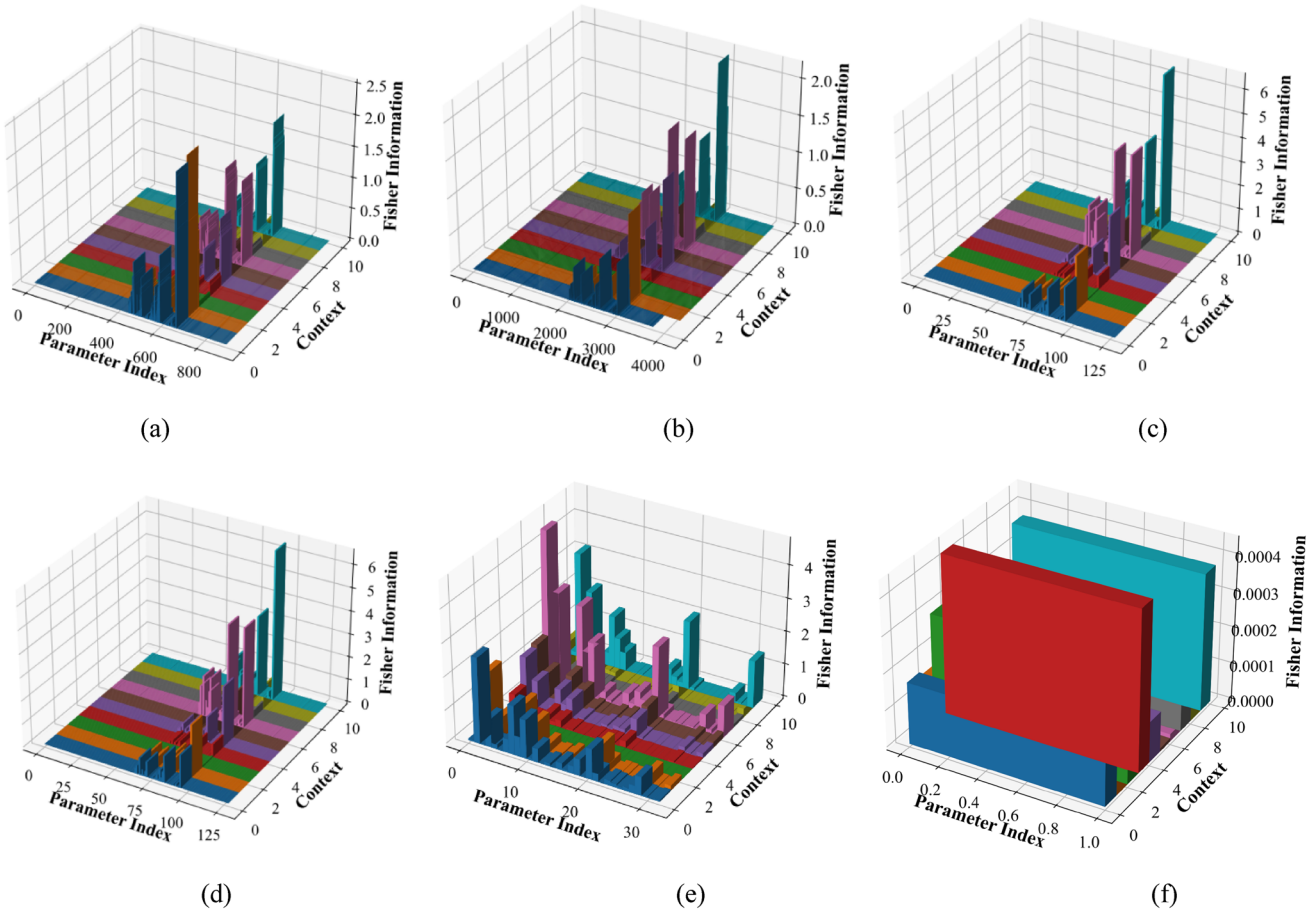


Fig. 10 Fisher information changes for different parameters during the training of influenza data by CEL, **a** lstm.weight_ih_10, **b** lstm.weight_hh_10, **c** lstm.bias_ih_10, **d** lstm.bias_hh_10, **e** linear.weight, and **f** linear.bias

thoroughly examine FIM's significance, indicating its impact on various facets of the model and its broader implications in addressing the challenges of Domain-IL.

In our proposed model, several critical elements drive its functionality. These encompass parameters are presented in Table 6. These parameters, characterized by their unique shapes, undergo iterative adjustments during training to minimize prediction errors, and the FIM serves as a guiding compass, unveiling which elements exhibit heightened responsiveness to changes when confronted with different disease datasets. Those with higher FIM values are better equipped to assimilate new information; thus, the model allocates maximum focus.

Similary, EWC prevents the model from erasing previously acquired knowledge as it learns new information. FIM is critical in determining the protection level required for each component through EWC. It introduces a regularization term into the base loss, resulting in a regularized loss, as shown in Fig. 8 for all diseases.

This regularized loss is initially higher at the beginning of context training but gradually decreases with the introduction of new contexts. So, the model's performance becomes stable after the last context. Iterations also show high trends where standard deviation abruptly changes for contexts, for example, in the mpox for contexts 3 and 5. Components deemed more critical due to their higher FIM values receive the highest protection, preserving vital knowledge while permitting other facets of the model to evolve.

To visually represent the evolution of different model components, we employ 3D graphs in Figs. 9, 10, 11. These visualizations clearly show which components remain stable and which undergo transformation when faced with new disease data. The 3D plots map Fisher information across several contexts and parameter indices, providing insights into the shifting significance of LSTM parameters across varying contexts. As the model encounters an increasing number of contexts, the divergence in parameter influence becomes more distinct. Certain parameters rise in importance, signifying their adaptability across diverse contexts. In contrast, some parameters, notably among bias terms like *lstm.bias_ih_l0* and *lstm.bias_hh_l0,* appear to diminish in

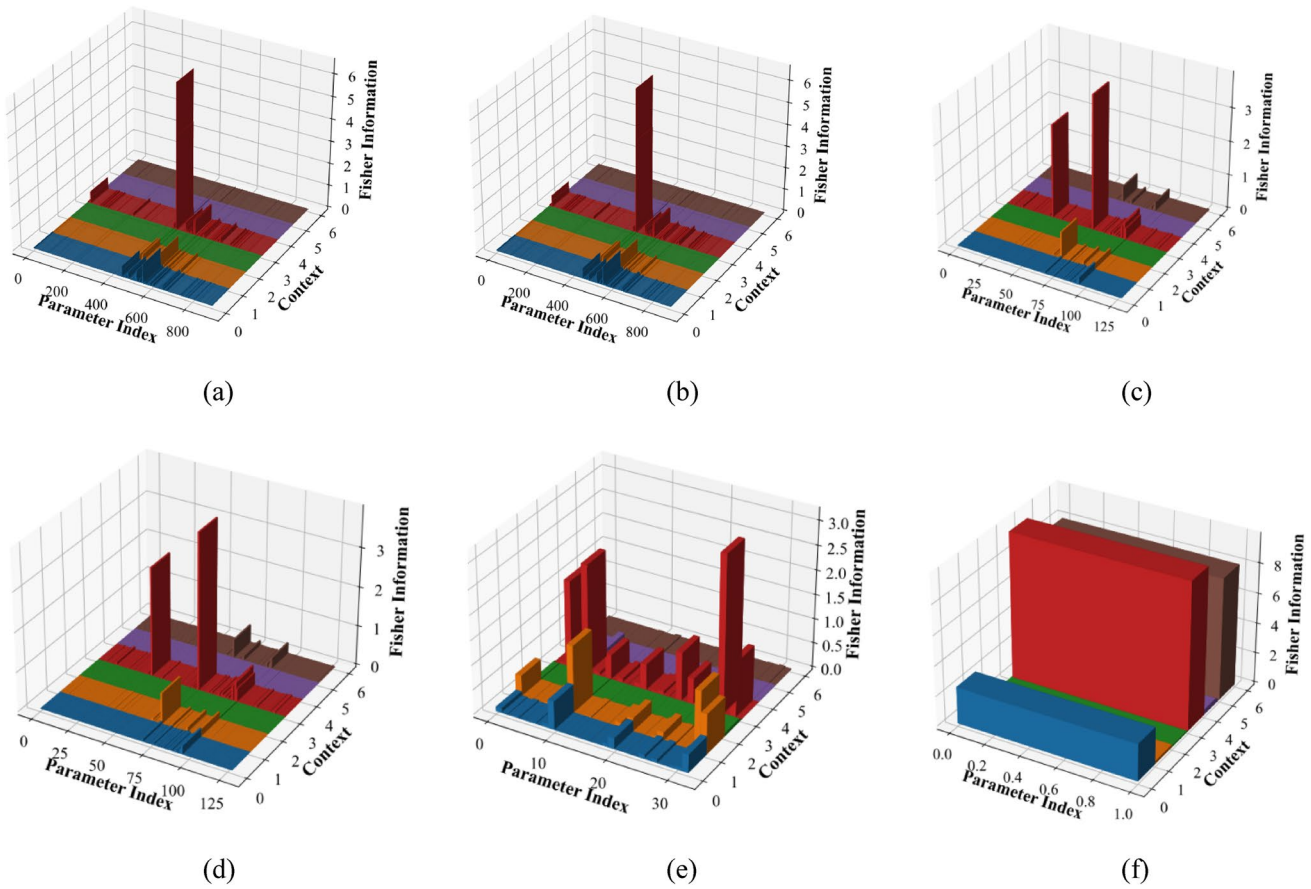


Fig. 11 Fisher information changes for different parameters during the training of measles data by CEL, **a** lstm.weight_ih_10, **b** lstm.weight_hh_10, **c** lstm.bias_ih_10, **d** lstm.bias_hh_10, **e** linear.weight, and **f** linear.bias

Table 7 Proposed CEL's performances across different numbers of contexts for disease (D) (mpox (Mx), influenza (I), measles (M)) and comparison with GEM [42], XdG [47], and EMC [41]

Models	D	Number of Contexts				
		6	7	8	9	10
GEM	Mx	0.722	0.698	0.355	0.712	0.805
XdG		0.601	0.584	0.318	0.689	0.751
EMC		0.634	0.720	0.325	0.702	0.780
CEL		0.728	0.729	0.358	0.718	0.818
GEM	I	0.844	0.817	0.809	0.861	0.887
XdG		0.711	0.713	0.785	0.754	0.800
EMC		0.775	0.742	0.812	0.786	0.776
CEL		0.847	0.817	0.827	0.867	0.887
GEM	M	0.555	0.022	0.444	0.477	0.475
XdG		0.451	0.021	0.413	0.388	0.307
EMC		0.473	0.010	0.423	0.405	0.336
CEL		0.657	0.029	0.457	0.498	0.577

influence, hinting at potential redundancies. This observation underscores that not all parameters hold equal significance across all contexts.

5.4 Segmentation Strategy Impact

One significant factor guiding our segmentation strategy is the average *R*-squared values, which offer a valuable metric for understanding the impact of different context divisions

on model performance. We opted for a finer segmentation strategy for the mpox (Mx) and influenza (I) outbreaks, which exhibited relatively higher average *R*-squared values across the 10th context, presented with bold values in Table 7. This choice was based on the observation that more contexts allowed our models to capture the nuances in the data distribution, resulting in improved predictive performance. Conversely, for Measles (M), which displayed a much lower *R*-squared with a higher number of contexts (with minimum on context 7 (0.01) to max 0.577 on 10, which is even lower than 60%), a coarser segmentation into 6th contexts (highlighted with bold values) was preferred as it is giving 0.657 value. This segmentation choice ensured that we did not unnecessarily burden the models with an excessive number of contexts while still enabling effective learning. Determining the optimal number of contexts (*N*) relies on some principles. Observing changes in memory requirements with varying *N* strikes a balance between computational efficiency and model accuracy. In addition, leveraging more historical data is crucial for effectively capturing underlying patterns. Furthermore, this ensures that the segmentation strategy aligns with data characteristics and practical computational considerations, contributing to an optimized strategy for enhanced model performance across diverse outbreaks.

6 Conclusion and Future Work

In this study, we introduced and evaluated the CEL model as a novel approach to continual learning that synergistically combines EWC with neural networks. The primary motivation behind CEL is to address the catastrophic forgetting phenomenon of DNNs, which is a significant challenge in continual learning, especially in disease outbreak prediction.

Our extensive experiments across three distinct diseases (mpox, influenza, and measles) provided valuable insights into the efficacy of the CEL model. The results consistently demonstrated that CEL offered a robust and stable performance across varying contexts, outpacing other state-of-the-art continual learning models (GEM, XdG, and EMC) and traditional models (LSTM and Transformer). Notably, the proposed CEL model's ability to maintain high *R*-squared values in evaluation and reevaluation with 65% minimal forgetting and 18% higher memory stability (Table 5) highlighted its potential as a reliable model for disease prediction. The consistent performance of CEL across different diseases highlights its versatility and adaptability. By leveraging the memory retention capabilities of EWC and the sequential data processing strengths of LSTM, our proposed model offers a balanced approach that captures both the temporal patterns and the evolving nature of disease data.

However, our CEL model exhibits specific technical weaknesses and areas for future improvement. Firstly, despite the valuable insights provided by *R*-squared utilized in assessing the regression model, exclusive reliance on this metric is cautioned against. The inclusion of RMSE is recommended. Secondly, CEL's effectiveness may vary depending on the characteristics of disease data, such as when diseases demonstrate irregular or unpredictable patterns. Lastly, the model might face challenges when confronted with abrupt and significantly different context changes, such as emerging diseases with distinct patterns.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12539-024-00675-2.

Acknowledgements This work is supported by National Natural Science Foundation of China (62273322), Shenzhen Medical Research Fund (D2401005), Natural Science Foundation of Guangdong Province (2024A1515012269), and Shenzhen Science and Technology Program (CJGJZD20220517142000002).

Author Contributions Saba Aslam: Data curation, Methodology, Software, Writing—original draft, Writing—review & editing, Visualization. Abdur Rasool: Writing—review & editing, Visualization, Formal analysis, Validation. Xiaoli Li: Formal analysis, Visualization. Hongyan Wu: Conceptualization, Funding, Methodology, Supervision, Resources, Formal analysis, Project administration. All authors have read and agreed to the published version of the manuscript.

Data Availability The data for this work is available at https://github.com/aI-area/CEL.

Declarations

Conflict of Interest The authors declare no conflict of interest.

References

1. Singh R, Singh R (2023) Applications of sentiment analysis and machine learning techniques in disease outbreak prediction – a review. Mater Today Proc 81:1006–1011. https://doi.org/10.1016/j.matpr.2021.04.356
2. Zhang J, Zhou P, Zheng Y (2023) Predicting influenza with pandemic-awareness via dynamic virtual graph significance networks. Comput Biol Med 158:106807. https://doi.org/10.1016/j.compbiomed.2023.106807
3. Feng D, Chen H (2021) A small samples training framework for deep learning-based automatic information extraction: case study of construction accident news reports analysis. Adv Eng Inform 47:101256. https://doi.org/10.1016/j.aei.2021.101256
4. Hadsell R, Rao D, Rusu AA et al (2020) Embracing change: continual learning in deep neural networks. Trends Cogn Sci 24(12):1028–1040. https://doi.org/10.1016/j.tics.2020.09.004
5. Kudithipudi D, Neftci E, Sandamirskaya Y et al (2022) Biological underpinnings for lifelong learning machines. Nat Mach Intell 4(3):196–210. https://doi.org/10.1038/s42256-022-00452-0
6. Chen Z, Liu B (2018) Lifelong machine learning, 2nd edn. Springer, Cham. https://doi.org/10.1007/978-3-031-01581-6
7. Parisi GI, Kemker R, Part JL et al (2019) Continual lifelong learning with neural networks: a review. Neural Netw 113:54–71. https://doi.org/10.1016/j.neunet.2019.01.012

8. Jenkins DA, Martin GP, Sperrin M et al (2021) Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? Diagn Progn Res 5(1):1–7. https://doi.org/10.1186/s41512-020-00090-3
9. van de Ven GM, Tuytelaars T, Tolias AS (2022) Three types of incremental learning. Nat Mach Intell 4(12):1185–1197. https://doi.org/10.1038/s42256-022-00568-3
10. Lee CS, Lee AY (2020) Clinical applications of continual learning machine learning. Lancet Digit Health 2(6):E279-e281. https://doi.org/10.1016/S2589-7500(20)30102-3
11. Perkonigg M, Waldstein SM, Prayer D et al (2021) Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nat Commun 12:5678. https://doi.org/10.1038/s41467-021-25858-z
12. Verwimp E, De Lange M, Masana M et al (2023) CLAD: a realistic continual learning benchmark for autonomous driving. Neural Netw 161:659–669. https://doi.org/10.1016/j.neunet.2023.02.001
13. Doshi K, Yilmaz Y (2020) Continual learning for anomaly detection in surveillance videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 254–255. https://api.semanticscholar.org/CorpusID:215814525
14. Devlin J, Chang MW, Lee K (2019) BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the NAACL, pp 4171–4186. https://doi.org/10.48550/arXiv.1810.04805
15. Lan Z, Chen M, Goodman S et al (2020) ALBERT: a lite BERT for self-supervised learning of language representations. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1909.11942
16. Liu Y, Ott M, Goyal N et al (2019) RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. https://doi.org/10.48550/arXiv.1907.11692
17. Radford A, Wu J, Child R et al (2019) Language models are unsupervised multitask learners. OpenAI Blog 1(8):9
18. Ke Z, Shao Y, Lin H (2022) Continual pre-training of language models. In: Proceedings of the Eleventh International Conference on Learning Representations. https://dblp.org/rec/conf/iclr/KeSLKK023.bib
19. Wu T, Caccia M, Li Z et al (2021) Pretrained language model in continual learning: a comparative study. In: International Conference on Learning Representations. https://doi.org/10.1007/s12539-024-00648-5
20. Rolnick D, Ahuja A, Schwarz J et al (2019) Experience replay for continual learning. Adv Neural Inf Process Syst 32. https://arxiv.org/pdf/1811.11682
21. Ramapuram J, Gregorova M, Kalousis A (2020) Lifelong generative modeling. Neurocomputing 404:381–400. https://doi.org/10.1016/j.neucom.2020.02.115
22. Aljundi R, Lin M, Goujaud B et al (2019) Gradient based sample selection for online continual learning. Adv Neural Inf Process Syst 32. https://api.semanticscholar.org/CorpusID:195345359
23. Zhang J, Hong Z, Chen Y (2020) Class-incremental learning via deep model consolidation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp 1131–1140. https://api.semanticscholar.org/CorpusID:83458730
24. Lopez-Paz D, Ranzato M (2017) Gradient episodic memory for continual learning. Adv Neural Inf Process Syst 30. https://proceedings.neurips.cc/paper/2017/file/f87522788a2be2d171666752f97ddebb-Paper.pdf
25. De Lange M, Aljundi R, Masana M et al (2021) A continual learning survey: defying forgetting in classification tasks. IEEE Trans Pattern Anal Mach Intell 44(7):3366–3385. https://doi.org/10.1109/TPAMI.2021.3057446
26. Mai Z, Li R, Jeong J et al (2022) Online continual learning in image classification: an empirical survey. Neurocomputing 469:28–51. https://doi.org/10.1016/j.neucom.2021.10.021
27. Masana M, Liu X, Twardowski B et al (2022) Class-incremental learning: survey and performance evaluation on image classification. IEEE Trans Pattern Anal Mach Intell 45(5):5513–5533. https://doi.org/10.1109/TPAMI.2022.3213473
28. Mirza MJ, Masana M, Possegger H (2022) An efficient domain-incremental learning approach to drive in all weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3001–3011. https://doi.org/10.1109/cvprw56347.2022.00339
29. Wang R, Ji C, Jiang Z et al (2021) A short-term prediction model at the early stage of the COVID-19 pandemic based on multi-source urban data. IEEE Trans Comput Soc Syst 8(4):938–945. https://doi.org/10.1109/tcss.2021.3060952
30. Aslam S, Aslam H, Manzoor A et al (2024) AntiPhishStack: LSTM-based stacked generalization model for optimized phishing URL detection. Symmetry 16(2):248. https://doi.org/10.3390/sym16020248
31. Kara A (2021) Multi-step influenza outbreak forecasting using deep LSTM network and genetic algorithm. Expert Syst Appl 180:115153. https://doi.org/10.1016/j.eswa.2021.115153
32. Yang S, Bao Y (2021) Comprehensive learning particle swarm optimization enabled modeling framework for multi-step-ahead influenza prediction. Appl Soft Comput 113:107994. https://doi.org/10.1016/j.asoc.2021.107994
33. Taylor SJ, Letham B (2018) Forecasting at scale. Am Stat 72(1):37–45. https://doi.org/10.1080/00031305.2017.1380080
34. Jha BK, Pande S (2021) Time series forecasting model for supermarket sales using FB-Prophet. In: Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), pp 547–554. https://doi.org/10.1109/iccmc51019.2021.9418033
35. Aslam S, Rasool A, Jiang Q (2021) LSTM-based model for real-time stock market prediction on unexpected incidents. In: Proceedings of the IEEE International Conference on Real-time Computing and Robotics (RCAR), pp 1149–1153. https://doi.org/10.1109/RCAR52367.2021.9517625
36. Pai PF, Liu CH (2018) Predicting vehicle sales by sentiment analysis of Twitter data and stock market values. IEEE Access 6:57655–57662. https://doi.org/10.1109/access.2018.2873730
37. Han L, Liang H, Chen H et al (2021) Convective precipitation nowcasting using U-Net model. IEEE Trans Geosci Remote Sens 60:1–8. https://doi.org/10.1109/tgrs.2021.3100847
38. Zhang F, Wang X, Guan J (2021) A novel multi-input multi-output recurrent neural network based on multimodal fusion and spatiotemporal prediction for 0–4 hour precipitation nowcasting. Atmosphere 12(12):1596. https://doi.org/10.3390/atmos12121596
39. Prasad GLV, Teja BR, Govathoti S (2023) Leveraging ARMA and ARMAX time-series forecasting models for rainfall prediction. In: Proceedings of the 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS), vol 1, pp 353–357. https://doi.org/10.1109/icaccs57279.2023.10113031
40. Kirkpatrick J, Pascanu R, Rabinowitz N et al (2017) Overcoming catastrophic forgetting in neural networks. Proc Natl Acad Sci USA 114(13):3521–3526. https://doi.org/10.1073/pnas.1611835114
41. Schillaci G, Schmidt U, Miranda L (2021) Prediction error-driven memory consolidation for continual learning: on the case of adaptive greenhouse models. KI-Künstliche Intelligenz 35:71–80. https://doi.org/10.1007/s13218-020-00700-8
42. Amalapuram SK, Tadwai A, Vinta R (2022) Continual learning for anomaly-based network intrusion detection. In: Proceedings of the 2022 14th International Conference on Communication

Systems & Networks (COMSNETS), pp 497–505. https://doi.org/10.1109/comsnets53615.2022.9668482
43. Rebuffi SA, Kolesnikov A, Sperl G (2017) iCaRL: incremental classifier and representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 2001–2010. https://doi.org/10.1109/CVPR.2017.587
44. Shin H, Lee JK, Kim J et al (2017) Continual learning with deep generative replay. Adv Neural Inf Process Syst 30. https://arxiv.org/pdf/1705.08690
45. Li Z, Hoiem D (2017) Learning without forgetting. IEEE Trans Pattern Anal Mach Intell 40(12):2935–2947. https://doi.org/10.1109/TPAMI.2017.2773081
46. Aljundi R, Babiloni F, Elhoseiny M (2018) Memory aware synapses: learning what (not) to forget. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 139–154. https://doi.org/10.1007/978-3-030-01219-9_9
47. Masse NY, Grant GD, Freedman DJ (2018) Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. Proc Natl Acad Sci USA 115(44):E10467–E10475. https://doi.org/10.1073/pnas.1803839115
48. Rajasegaran J, Hayat M, Khan SH et al (2019) Random path selection for continual learning. Adv Neural Inf Process Syst 32. https://papers.nips.cc/paper/9429-random-path-selection-for-continual-learning
49. Xu J, Zhu Z (2018) Reinforced continual learning. Adv Neural Inf Process Syst 31. https://arxiv.org/pdf/1805.12369
50. Mallya A, Lazebnik S (2018) PackNet: adding multiple tasks to a single network by iterative pruning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7765–7773. https://doi.org/10.1109/CVPR.2018.00810
51. Masana M, Tuytelaars T, Van de Weijer J (2021) Ternary feature masks: zero-forgetting for task-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3570–3579. https://doi.org/10.1109/cvprw53098.2021.00396
52. Chaudhry A, Dokania PK, Ajanthan T (2018) Riemannian walk for incremental learning: understanding forgetting and intransigence. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 532–547. https://arxiv.org/pdf/1801.10112
53. Aich A (2021) Elastic weight consolidation (EWC): Nuts and bolts. arXiv:2105.04093. https://doi.org/10.48550/arXiv.2105.04093
54. Ly A, Marsman M, Verhagen J et al (2017) A tutorial on fisher information. J Math Psychol 80:40–55
55. Wang Z, Bovik AC (2009) Mean squared error: love it or leave it? A new look at signal fidelity measures. IEEE Signal Process Mag 26(1):98–117. https://doi.org/10.1109/MSP.2008.930649
56. Mathieu E, Spooner F, Dattani S et al (2022) Mpox (monkeypox). OurWorldInData.org. https://ourworldindata.org/monkeypox. Accessed 20 Aug 2024
57. Centers for Disease Control and Prevention, National Center for Immunization and Respiratory Diseases (NCIRD) FluView Interactive. https://www.cdc.gov/flu/weekly/fluviewinteractive.htm. Accessed 18 Aug 2024
58. European Centre for Disease Prevention and Control's Atlas Platform. https://atlas.ecdc.europa.eu/public/index.aspx?Dataset=27&HealthTopic=20. Accessed 18 Aug 2024
59. Vaswani A, Shazeer N, Parmar N (2017) Attention is all you need. Adv Neural Inf Process Syst 30. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf