Computers in Biology and Medicine 175 (2024) 108548

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Deep continual learning for medical call incidents text classification under the presence of dataset shifts

Pablo Ferri [a,*], Vincenzo Lomonaco [b], Lucia C. Passaro [b], Antonio Félix-De Castro [c], Purificación Sánchez-Cuesta [c], Carlos Sáez [a], Juan M. García-Gómez [a]

[a] Biomedical Data Science Laboratory (BDSLab), Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València (UPV), Valencia, Spain
[b] Department of Computer Science, University of Pisa (Unipi), Pisa, Italy
[c] Conselleria de Sanitat Universal i Salut Pública, Generalitat Valenciana (GVA), Valencia, Spain

ARTICLE INFO

Keywords:
Continual learning
Deep learning
Dataset shifts
Emergency medical call incidents
Emergency medical dispatch
Natural language processing

ABSTRACT

The aim of this work is to develop and evaluate a deep classifier that can effectively prioritize Emergency Medical Call Incidents (EMCI) according to their life-threatening level under the presence of dataset shifts. We utilized a dataset consisting of 1 982 746 independent EMCI instances obtained from the Health Services Department of the Region of Valencia (Spain), with a time span from 2009 to 2019 (excluding 2013). The dataset includes free text dispatcher observations recorded during the call, as well as a binary variable indicating whether the event was life-threatening. To evaluate the presence of dataset shifts, we examined prior probability shifts, covariate shifts, and concept shifts. Subsequently, we designed and implemented four deep Continual Learning (CL) strategies—cumulative learning, continual fine-tuning, experience replay, and synaptic intelligence—alongside three deep CL baselines—joint training, static approach, and single fine-tuning—based on DistilBERT models. Our results demonstrated evidence of prior probability shifts, covariate shifts, and concept shifts in the data. Applying CL techniques had a statistically significant ($\alpha = 0.05$) positive impact on both backward and forward knowledge transfer, as measured by the F1-score, compared to non-continual approaches. We can argue that the utilization of CL techniques in the context of EMCI is effective in adapting deep learning classifiers to changes in data distributions, thereby maintaining the stability of model performance over time. To our knowledge, this study represents the first exploration of a CL approach using real EMCI data.

1. Introduction

Emergency Medical Dispatch (EMD) involves the reception and handling of requests for medical assistance in emergency situations [1]. It is a challenging task characterized by a high level of uncertainty, limited decision time, and scarce resources [2]. Given the potential severe consequences, including patient mortality and significant costs, associated with errors in this critical environment, there is a need for decision support tools to enhance call-taking situations.

The EMD process consists of two main components: triage, which assesses the priority of incidents, and resource allocation, which assigns the most appropriate resources to respond to each incident. In the context of triage, dispatchers typically follow predefined clinical guidelines in the form of decision trees [2,3]. Examples of these triage protocols include the Emergency Severity Index [4] and the Manchester Triage System [5]. However, these clinical algorithms have two main limitations: firstly, they are based on archetypical cases, overlooking the vast number of incidents with complex characteristics, and secondly, they heavily rely on structured clinical information, which is not always available during emergency medical calls. As a result, these algorithms are unable to automatically handle unstructured data, such as free text.

During emergency medical calls, a significant amount of data is generated [6]. While these data are typically stored in health institution databases, they are often underutilized, only used for basic business intelligence analyses. Consequently, the latent information contained within these data, including hidden statistical patterns, is not considered to improve triage protocols. Moreover, a substantial portion of this data is in the form of unstructured information, which cannot be automatically processed by current triage protocols [7,8]. Therefore,

* Corresponding author.
E-mail addresses: pabferb2@upv.es (P. Ferri), vincenzo.lomonaco@unipi.it (V. Lomonaco), lucia.passaro@unipi.it (L.C. Passaro), felix_antdec@gva.es (A. Félix-De Castro), sanchez_pur@gva.es (P. Sánchez-Cuesta), carsaesi@upv.es (C. Sáez), juanmig@ibime.upv.es (J.M. García-Gómez).

https://doi.org/10.1016/j.compbiomed.2024.108548
Received 8 September 2023; Received in revised form 11 April 2024; Accepted 28 April 2024
Available online 1 May 2024

an alternative approach is needed to complement the limitations of existing triage protocols and enhance EMD processes.

Machine Learning (ML) stands out as one of the most promising approaches in the EMD environment, given its ability to autonomously identify meaningful patterns within datasets. Among ML methodologies, Deep Learning stands out due to its exceptional capacity for conducting feature extraction automatically from complex data types [9]. This capability is further enhanced by advancements in training techniques, which predominantly rely on gradient-based methods. These methods can be augmented with evolutionary computing techniques for either direct training purposes or to aid in the optimization of hyperparameters [10]. Deep Learning is particularly adept at processing unstructured data, such as free text, which is prevalent in EMD databases. Furthermore, it represents the cutting-edge in numerous intricate tasks. For instance, in object recognition, Convolutional Neural Networks (CNNs) [11] have become the standard model. Sentiment analysis frequently utilizes Transformer architectures [12], while text generation has seen significant contributions from generative models, including Text Generative Adversarial Networks (TextGANs) [13] and Transformer-based architectures.

Numerous studies have demonstrated the value offered by ML tools in EMD domain. For instance, [14] developed ML-based models to predict the risk associated with individual patients in prehospital emergency medical events. Their findings revealed that ML-based scores surpassed rule-based triage algorithms and human prioritization decisions in terms of performance. Similarly, [15] explored the application of ML in detecting cardiac arrest from audio files of emergency calls. They demonstrated that ML techniques can increase sensitivity in cardiac arrest detection while maintaining a reasonable level of specificity. Furthermore, [16] conducted an evaluation of different ML models and their impact on the early detection of under-triaged patients. Their study revealed that ML models can effectively aid in identifying under-triaged patients, leading to improved patient outcomes.

In the specific context of the emergency medical services of the Valencian Region, Spain, a project was undertaken with the objective of developing ML models utilizing historical EMD data for predicting incident priority and assessing its influence on the EMD process. As part of this project, a deep ensemble multitask deep learning model called DeepEMC^2 was created, as described by [17]. The model showed improvements in performance metrics compared to the existing in-house triage system. Specifically, it achieved better predictions in terms of life-threatening (+12.5%), admissible response delay (+17.5%), and emergency system jurisdiction (+5.1%). Notably, the model's success was attributed mainly to features extracted from free text, which proved to be more predictive than the clinical variables recorded during the call.

However, it should be noted that the data used to train the DeepEMC^2 model only covered the period from 2009 to 2012. As information systems, dispatchers, coordination centers, and demographics evolve over time, dataset shifts occur, leading to changes in the joint probability distribution of inputs and outputs between the training and testing stages [18,19]. In fact, the Valencian EMCI information system underwent significant changes in 2013, impacting the registration, organization, and storage of incident information. Furthermore, the in-house dispatch triage protocol underwent substantial modifications, influencing how incidents were triaged, and there were notable changes in dispatcher personnel. Additionally, despite their relatively small scale, other changes occurred within the time frame of our data, including adjustments to dispatcher training programs that affected incident handling. Minor updates to the in-house triage protocol and dispatcher personnel also took place. Fig. 1 provides a visual representation of these changes. Consequently, the DeepEMC^2 model developed in [17] using data from 2009 to 2012 may require adjustments to mitigate potential performance degradation resulting from distributional shifts caused by these changes.

Therefore, it is reasonable to consider the incorporation of Continual Learning (CL) strategies to address the challenge of dataset shifts in EMCI. CL strategies facilitate the integration of new knowledge while avoiding catastrophic forgetting [20,21], enabling a sustainable learning process over time and providing adaptable decision support for call-takers. To implement this CL approach, we exploit multiple learning experiences—i.e., a set composed of data samples that can be used to update a ML model—within our EMCI data, each associated with a different time period or batch of data [22]. Consequently, multiple data streams were derived from each batch, and the deep models learn from these streams according to the CL strategy defined.

Previous studies on Valencia EMCI data have emphasized the significance of free text in categorizing incident severity [8,17]. Given that dataset shifts are inherent in our emergency medical domain, as illustrated in Fig. 1, any ML text model intended for deployment in supporting emergency medical triage must possess the capability to adapt to these changes. Considering that CL provides a state-of-the-art approach to this problem, we find it reasonable to adopt this approach to ensure consistent decision support over time.

The primary objective of this work is to investigate the extent to which various CL strategies enable lifelong adaptation of deep triage models over time. While acknowledging the inevitable negative impact on performance due to changes in data distributions in real-world scenarios, our framework aims to minimize such effects by leveraging CL techniques. To achieve this, we first assess the presence of dataset shifts and subsequently, explore and evaluate multiple CL pipelines designed to mitigate the adverse effects on model performance resulting from distributional drifts.

Prior research has primarily focused on developing ML models tailored for handling medical data in the presence of temporal distributional drifts. In their work [23], authors advocated for the construction of robust models by estimating invariant properties across time, integrating this approach with unsupervised domain adaptation. Another study [24] addressed the issue of temporal distributional shift, employing parsimonious models generated through specific feature selection methods. In [25], the authors proposed incorporating pretrained foundational models into their developments, akin to BioBert [26], albeit with a focus on Electronic Health Records. Similarly, in [27], a patient re-weighting scheme is suggested as a strategy to alleviate the adverse effects of temporal dataset shifts.

The findings of our study contribute to the advancement of decision support systems in emergency medical triage, with practical applications in real settings. These systems have the potential to positively impact patient well-being and enhance the sustainability of health services. While the studies mentioned earlier concentrate on constructing ML models for managing medical data in the presence of temporal distributional drifts, this study marks the first attempt, to our knowledge, to address actual EMCI data through a CL approach. This represents a noteworthy contribution to the field and stands as one of the initial real-world applications of CL methods.

2. Materials

2.1. Dataset

A dataset comprising a total of 1 982 746 independent Emergency Medical Call Incidents (EMCI) was compiled from the Health Services Department (HSD) of the Valencian Region, covering the period from 2009 to 2019, with the exception of 2013 due to unavailability of data during the system update of the Valencian EMCI's information system. The use of this data was approved by the Institutional Review Board of the Health Services Department, and all necessary precautions were taken to ensure patient identity confidentiality.

The EMCI data consisted of both during-call and after-call information. During-call data were collected in real-time during the emergency medical call and included free text dispatcher observations written

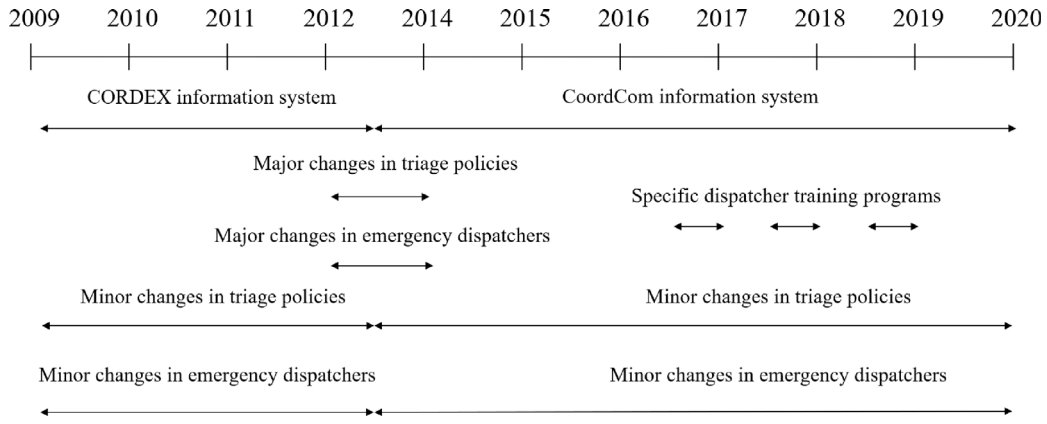

Fig. 1. Changes occurred in the Valencian emergency dispatch service within the time frame of our data.

Table 1
Examples of free text notes belonging to the dataset. The Life-threat column indicates whether the situation is life-threatening (1) or not (0).

Text	Life-threat
83-year-old woman with respiratory and cardiac insufficiency. Neoplastic disease in progression.	1
14 year old male with fever of 39°, he has been like this for 1 h and also general malaise.	0
85 year old woman with a lot of fatigue and cough since yesterday. Today saturation at 85, she is on oxygen at home.	0

Table 2
Basic descriptive statistics. The number of instances for each class—life-threatening and non-life-threatening—by year is detailed, as well as the median number of words per class and year.

Year	Number of instances		Median number of words	
	Life-threat	Non-life-threat	Life-threat	Non-life-threat
2009	71 064	111 179	10	10
2010	71 707	106 618	11	10
2011	74 641	105 993	11	11
2012	75 740	104 713	11	11
2014	61 183	104 207	9	12
2015	68 641	120 810	9	12
2016	70 326	127 450	8	11
2017	79 310	158 064	9	11
2018	79 318	152 448	9	12
2019	82 318	157 016	10	12

in the Spanish language. These observations were short sentences describing the incident, such as “stabbing chest pain with shortness of breath”, “fever, general malaise, vomiting” or “traffic accident, profuse bleeding, unconscious”. During inference, these observations were used as input for prediction.

After-call data were recorded at a later time, following the completion of the call. This data encompassed information such as physician diagnosis, hospitalizations, urgent care visits, medical procedures, and treatments received by the patient. Importantly, these after-call data were not used during prediction but rather offline, for inferring whether the emergency event constituted a life-threatening situation, considering a mapping developed by expert physicians from the HSD of the Valencian Region. This binary variable served as the classification label in our work.

We provide in Table 1, some examples of the data considered to train and evaluate our deep triage models.

In Table 2, we present basic descriptive statistics, including the number of instances per class and year, as well as the median number of words per class and year.

2.2. Framework

Our experiments were implemented in Python [28], utilizing the libraries Numpy [29] and Pandas [30] for data management, PyTorch [31] and HuggingFace’s Transformers [32] for modeling, Avalanche [22] for continual learning, and Optuna [33] for hyperparameter tuning. The source code developed to carry out the dataset shift analyses and implement the different deep text continual pipelines can be accessed via the following link: https://github.com/bdslab-upv/DeepConText112.git.

3. Methods

3.1. Data preparation

We utilized Natural Language Processing (NLP) techniques for both data preprocessing and inference with respect to free text variables. These techniques involved the utilization of language models, which are described in the next section. We applied pre-processing functions (e.g., lowercasing, removal of special characters, and accent marks) to enhance the language model encoding capability. Additionally, we employed sub-word tokenization using WordPiece [34] to reduce the size of the vocabulary. We selected WordPiece as it is the default tokenizer for DistilBERT models [35]—which we will focus on in our study as further explained in the Modeling section—and also for BERT models [36] from which DistilBERT is derived. Additionally, its efficiency in managing out-of-vocabulary words and enhancing generalization—owing to its reliance on sub-words—further motivated our choice. Subsequently, these sub-words were mapped to indexes, and padding and truncation operations were applied to ensure that all text records shared the same sequence length, facilitating computation. Boolean attention masks were generated to exclude the impact of padding indexes.

The data were organized into ten learning experiences, each representing one year. Each learning experience consisted of a training stream and a test stream. Similarly, each training stream comprised a pure training stream and a validation stream, which were used for hyperparameter tuning operations without overfitting to the test stream. Next, the data arrangement process is presented in Table 3.

3.2. Dataset shifts assessment

In the context of a classification problem, as described in [19], where we have input features denoted as x and an output variable denoted as y, a dataset shift occurs when the joint probability distri-

Table 3
Data arrangement process. The data is divided into experiences, with each experience corresponding to a different year. Furthermore, each experience consists of three distinct and non-overlapping data streams: a pure training stream, a validation stream, and a test stream. Abbreviations: Exp, Experience.

Exp	Year	Pure train	Validation	Test	Total
1	2009	101 669	43 710	36 864	182 243
2	2010	100 147	42 465	35 713	178 325
3	2011	101 245	43 459	35 930	180 634
4	2012	101 253	43 399	35 801	180 453
5	2014	92 396	39 860	33 134	165 390
6	2015	106 013	45 461	37 977	189 451
7	2016	110 883	47 302	39 591	197 776
8	2017	132 934	56 986	47 454	237 374
9	2018	129 963	55 692	46 111	231 766
10	2019	133 835	57 525	47 974	239 334

butions of the training and test data differ:

$$P_{train}(x, y) \neq P_{test}(x, y) \tag{1}$$

This variation can be attributed to three different sources of drift: prior probability shifts, covariate shifts, and concept shifts [19]. In the following sections, we provide a concise explanation of each type of shift and how it was meticulously examined in our dataset:

3.2.1. Prior probability shift

A prior probability shift arises when there is a difference between the class variable distributions in the training and testing datasets:

$$P_{train}(y) \neq P_{test}(y) \tag{2}$$

To assess the presence of prior probability shifts in our study, we calculated the empirical probabilities of life-threatening events over time. These probabilities were plotted on a temporal graph, and a stationarity test was conducted. Specifically, we employed the Kwiatkowski–Phillips–Schmidt–Shin test [37], which tests the null hypothesis of data distribution stationarity.

3.2.2. Covariate shift

Covariate shift refers to changes in the distribution of input features:

$$P_{train}(x) \neq P_{test}(x) \tag{3}$$

To evaluate the presence of distributional changes in these covariates, we constructed multiple pairwise text classification models (refer to the Modeling section for more information about model structure and rationale) for each pair of years. Subsequently, we compared their performance in terms of the Area Under Curve (AUC) against the expected performance of a random model. This comparison allowed us to determine whether the model could predict the year from which the data originated. If the model demonstrated the ability to make accurate predictions, it would indicate the presence of a covariate shift, suggesting that the covariates varied across different years.

3.2.3. Concept shift

Concept shift occurs when there is a variation in the conditional probability of the outcome with respect to the input features between different sets:

$$P_{train}(y|x) \neq P_{test}(y|x) \tag{4}$$

To assess the presence of concept shift, we trained a deep model (refer to the Modeling section for more details about the model architecture and rationale) for each year and evaluated its performance across all years. Any differences in model performance among the years could be interpreted as variations resulting from changes in the conditional distribution.

3.3. Modeling

The emergence of new deep learning architectures, such as Transformers [12], has significantly improved the performance of a wide array of NLP tasks. The attention mechanism in Transformers enables the model to access information from all elements of a text, allowing for contextual modeling of word and sentence meanings. Additionally, the Transformer architecture is well-suited for transfer learning in NLP, where knowledge gained from a more general task is utilized to specialize a model for new problems with limited data. This transfer of linguistic knowledge is achieved through pre-training and fine-tuning. Pre-training involves training language models using unsupervised learning tasks on extensive collections of text data, while fine-tuning involves further adapting the pre-trained model to specific supervised learning tasks, such as sequence classification. Transformer-based architectures have achieved state-of-the-art results across a wide range of NLP tasks. To ensure the effectiveness of the model in the presence of dataset shifts, we combine this paradigm with continual learning (CL).

In the previous DeepEMC^2 model, deep models based on the BERT architecture [17,36] were employed, resulting in a significant improvement compared to non-deep learning approaches. However, considering that we are evaluating multiple CL strategies, the data volume is large, and our main objective is to study different CL pipelines, we opted to use the DistilBERT [35] model in this work. DistilBERT offers an excellent balance between performance and efficiency, as it has significantly fewer parameters than BERT with only a minor performance decrease due to knowledge distillation [38]. In addition, it can be used locally, eliminating the privacy risks that come into play when using APIs—such as those encountered with models like ChatGPT [39]—which is particularly prudent given that we are handling sensitive data. It is important to note that we did not train our DistilBERT models from scratch; instead, we utilized the pretrained version available at [32] and adopted a transfer learning approach by fine-tuning the model for the specific downstream task. Therefore, our model architecture consists of:

1. An embedding block, which includes a word embedding layer, positional encoding layer [12], layer normalization layer [40], and dropout layer [41].
2. Multiple Transformer blocks, each composed of multi-head self-attention layers, layer normalization layers, and feed-forward layers.
3. An output block consisting of feed-forward layers, with the last layer utilizing softmax as the activation function.

For parameter tuning, we utilized the AdamW [42] optimizer, a variant of the Adam [43] algorithm, known for its suitability in training Transformer models [42]. The model was trained using a mini-batch training approach, and the loss function employed was cross-entropy [44], weighted to address class imbalance:

$$L = \sum_{n=1}^{N} \frac{-\sum_{c=1}^{C} \frac{1}{v_c} log \frac{\exp^{x_{n,c}}}{\sum_{i=1}^{C} \exp^{x_{x,i}}} y_{n,c}}{N} \tag{5}$$

Here, N denotes the mini-batch size, C represents the number of classes, v indicates the class frequency in the dataset, x denotes the logits, and y refers to the true target value.

3.4. Continual learning baselines

To assess the added value of including CL strategies for model adaptation over time, we employed three baseline techniques: a static model, single fine-tuning, and joint training. These baselines allow us to evaluate the impact of different approaches on model performance over the course of learning experiences.

3.4.1. Static model

The static model represents the scenario in which the model is not retrained over time, providing insight into how performance may decline if no action is taken. In this approach, we fine-tuned our pretrained DistilBERT model using only the data from the first learning experience (i.e., the year 2009). This approach serves as a lower performance bound for forward transfer of knowledge [45] since it does not update the model with instances from recent experiences.

3.4.2. Single fine-tuning

The single fine-tuning strategy involves retraining the original model, pretrained DistilBERT, using data exclusively from the current learning experience. For subsequent learning experiences, the model weights are not retained, and the model is reinitialized with the pretrained DistilBERT weights. This approach provides a lower performance bound for backward transfer of knowledge [45] since it does not retain information from previous experiences, considering only the data from the current experience.

3.4.3. Joint training

To estimate the best performance achievable by any CL strategy, we employed the joint training approach. This strategy involves training our deep model using data from all learning experiences, incorporating data from all years. While this approach is not applicable in a real-world setting, as we do not have access to future data at a given year, implementing this approach allows us to establish an upper bound for performance in terms of both forward and backward transfer of knowledge.

3.5. Continual learning strategies

We evaluated the following continual learning strategies:

3.5.1. Cumulative

The cumulative strategy involves re-estimating model parameters using data from the current learning experience as well as all the data encountered in previous experiences. This approach utilizes all available information up to that point, but it can be computationally expensive and may not be applicable if data from certain time periods is not accessible due to privacy or regulatory concerns.

3.5.2. Continual fine-tuning

The continual fine-tuning strategy is based on an incremental fine-tuning process. At each learning experience (in our case, the year), the model weights are initialized with the weights from the previous experience. Training at the current experience only considers the data from that particular experience.

3.5.3. Experience replay

The experience replay strategy relies on an external memory, known as a replay buffer, with a predefined size B. This buffer stores data samples from previous learning experiences. At each experience, data samples are sampled from the replay buffer, allowing the model to retain information about previous data patterns. This approach does not require as much computational resources as the cumulative strategy.

3.5.4. Synaptic intelligence

Synaptic intelligence [46] is a regularization-based strategy that mitigates catastrophic forgetting by incorporating a knowledge retention penalty into the loss function. Unlike the previous strategies, it does not rely on resampling or storing data from all previous experiences. The loss function to optimize at experience e follows the structure:

$$L_e = H_e + c \sum_{k=1}^{K} \Omega_k^e (\tilde{\theta}_k - \theta_k)^2 \tag{6}$$

Here, H_e represents the standard loss to minimize at experience e (in our case, the per-class weighted cross-entropy loss), c is a global dimensionless weighting parameter, Ω_k^e is the per-parameter regularization strength for parameter k and experience e, $\tilde{\theta}_k$ denotes the value of parameter k at the previous experience, and θ_k represents the value of parameter k at the current learning experience.

3.6. Evaluation

3.6.1. Backward and forward knowledge transfer

To evaluate and compare the advantages and disadvantages of each CL strategy, as well as to assess their performance in comparison to the baseline techniques, we calculated their backward and forward transfer [45]. Backward transfer refers to how learning from a particular experience affects prior knowledge, while forward transfer refers to how learning from a specific experience influences the acquisition of future knowledge.

For each CL strategy l and experience e, we computed the backward and forward transfer using the following formulas:

$$BWT_e^l = \frac{1}{e-1} \sum_{i=1}^{e-1} M_i^l \tag{7}$$

$$FWT_e^l = \frac{1}{E-e} \sum_{i=e+1}^{E} M_i^l \tag{8}$$

Here, M_i^l represents the value of the performance metric of strategy l at experience i, and E denotes the total number of experiences (in our case, years).

Additionally, we calculated the global backward and forward transfer, which provides an average performance estimation across all experiences for a specific strategy:

$$BWT_{global}^l = \frac{1}{E-1} \sum_{j=1}^{E-1} BWT_j^l \tag{9}$$

$$FWT_{global}^l = \frac{1}{E-1} \sum_{j=1}^{E-1} FWT_j^l \tag{10}$$

Furthermore, we utilized multiple evaluation metrics to assess the performance of each strategy. Specifically, we obtained the Area Under the Curve (AUC), accuracy, recall, precision and F1-score.

Finally, 95% confidence intervals for the global backward and forward transfer were estimated for each strategy. To derive them, we followed the next expression:

$$CI_l^{95\%} = \bar{m} \pm 1.96 \frac{s_l}{\sqrt{E}} \tag{11}$$

Here $\bar{m}$ will correspond to the BWT_{global}^l or FWT_{global}^l, and s_l is the sample standard deviation computed with the series of BWT_e^l or FWT_e^l.

3.6.2. Impact on the health service department of the Valencian Region

To comprehensively evaluate and understand the potential contributions of applying our algorithm to the emergency medical triage process in the Valencian Region, we conducted a comparison between the accuracy of the in-house triage protocol and our model in determining the life-threatening level of incidents. This study focused on the top 5 most common severe incidents and the top 5 most common non-severe incidents, allowing us to analyze contributions based on incident types (life-threatening or not). Errors in determining severity in severe incidents indicate under-triage, while errors in non-severe incidents suggest over-triage. By assessing the model's comparative performance using the optimal CL pipeline against the performance of the protocol, we can discern the significant value our deep continual approach brings to the patient and the health system.

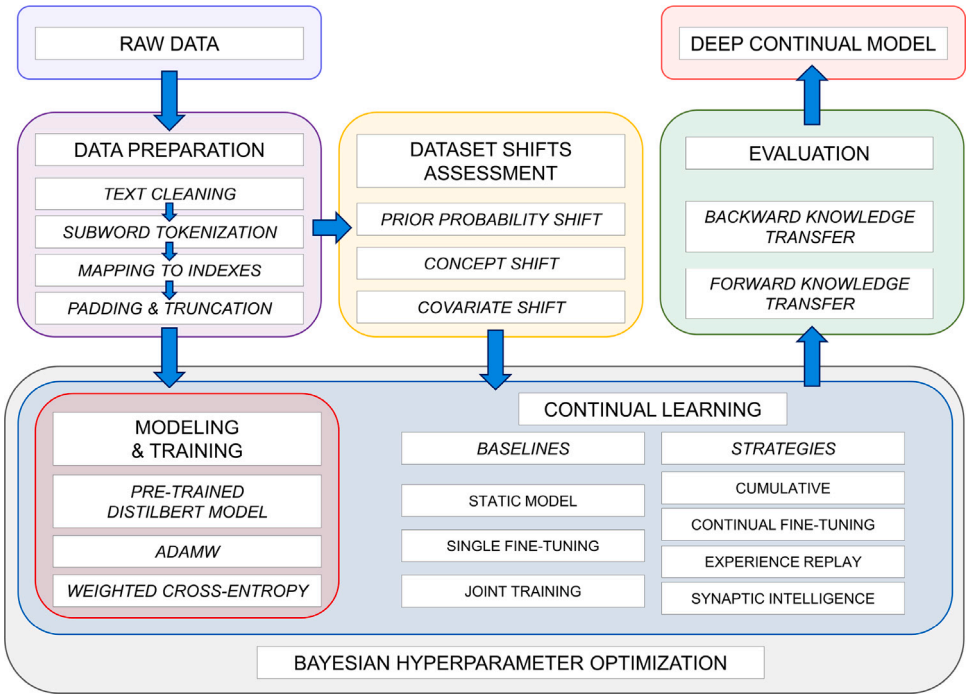

Fig. 2. Flow diagram of the different steps followed in our work.

3.7. Hyperparameter tuning

To determine the optimal hyperparameters, an automatic active learning approach [47] was employed. For each CL strategy, a set of hyperparameters was defined, including parameters such as learning rate and batch size. Additionally, a range of values was proposed for each hyperparameter. For example, for the learning rate, values of 0.0001 and 0.00001 were considered, while for the batch size, values of 16 and 32 were explored. The sampling space for the hyperparameters was discrete to avoid overfitting issues due to the curse of dimensionality.

A Bayesian optimization strategy was then employed, where an auxiliary probabilistic generative model was iteratively trained. The purpose of this model was twofold: (1) to estimate the probability of the objective performance metric (in this case, the weighted cross-entropy) given a specific set of hyperparameters, and (2) to sample new hyperparameter values on each iteration in the hope of improving the performance metric.

Once the optimal hyperparameters were determined through these experiments on the pure training and validation sets, they were used for the final retraining stage of each strategy. The models were retrained using the full training set, and the performance metrics reported in this work were obtained from the test set.

Next, we present a flow diagram in Fig. 2, summarizing the design and methods applied to our data.

4. Results

4.1. Dataset shifts assessment

4.1.1. Prior probability shift

The empirical probability of the life-threatening class over time is illustrated in Fig. 3:

The plot reveals two distinct drops in the class probability: one occurring between the years 2012 and 2014, and another between 2016 and 2017. However, from 2009 to 2012, the life-threatening class probability shows a gradual increase. Furthermore, the empirical probability appears to stabilize qualitatively in the remaining time periods, namely between 2014 and 2016, and between 2017 and 2019.

Table 4
P-value of the Kwiatkowski–Phillips–Schmidt–Shin test. Assessment of the stationarity of the empirical life-threatening probability distribution over time.

Kwiatkowski–Phillips–Schmidt–Shin test
.018*

In addition, the Kwiatkowski–Phillips–Schmidt–Shin test (Table 4) suggests rejecting the null hypothesis of stationarity.

The findings from Fig. 3 and Table 4 confirm the presence of a prior probability shift in our data.

4.1.2. Covariate shift

Fig. 4 presents the performance, in terms of AUC, of DistilBERT text classification models trained to predict the year from which the data originated. The models utilized the free text dispatcher observations as input features.

As shown in 4, the AUC values are consistently higher than those expected from a random model, which would have an AUC around 0.5. There is a clear distinction in the writing style of the free text fields between the 2009–2012 period and the 2014–2019 period, as indicated by the AUC of 1 on the test set. Moreover, within each time window, the AUC values gradually increase over time.

These observations confirm the presence of a distinct and abrupt covariate shift between the 2012 and 2014 periods, with smoother and gradual changes occurring within the 2009–2012 and 2014–2019 time windows.

4.1.3. Concept shift

The performance of the DistilBERT models trained to assess the presence of concept shift is depicted in Fig. 5:

As illustrated in Fig. 5, there is a significant performance drop in all models from 2009 to 2012, with the lowest F1-score observed in 2014. This confirms the presence of concept shift. Although there is a slight recovery in performance from 2015 to 2019, it still remains far from the values observed in the first period. Furthermore, the models trained on the 2015–2019 data show consistent performance within that time window, but experience a notable performance drop in the year 2014.

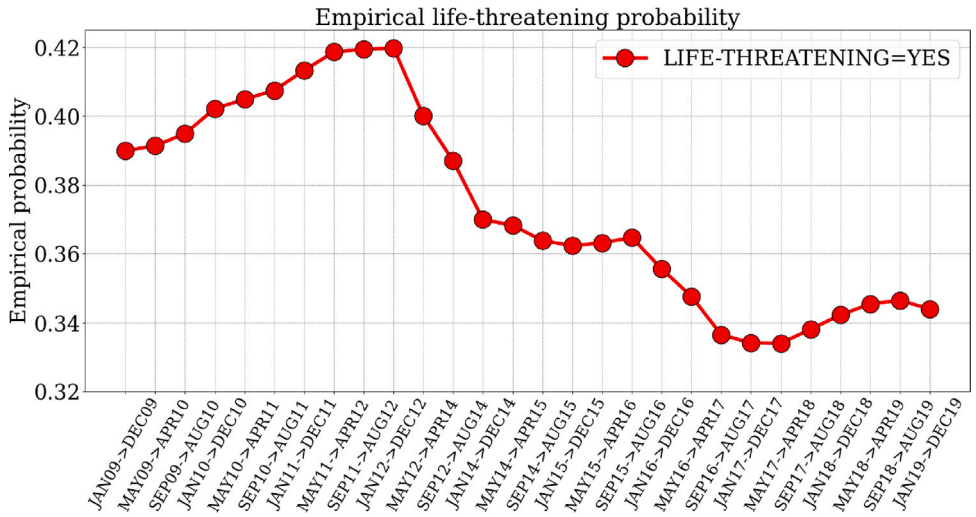

Fig. 3. Empirical life-threatening probability over time. A significant drop in the class probability is observed over time.

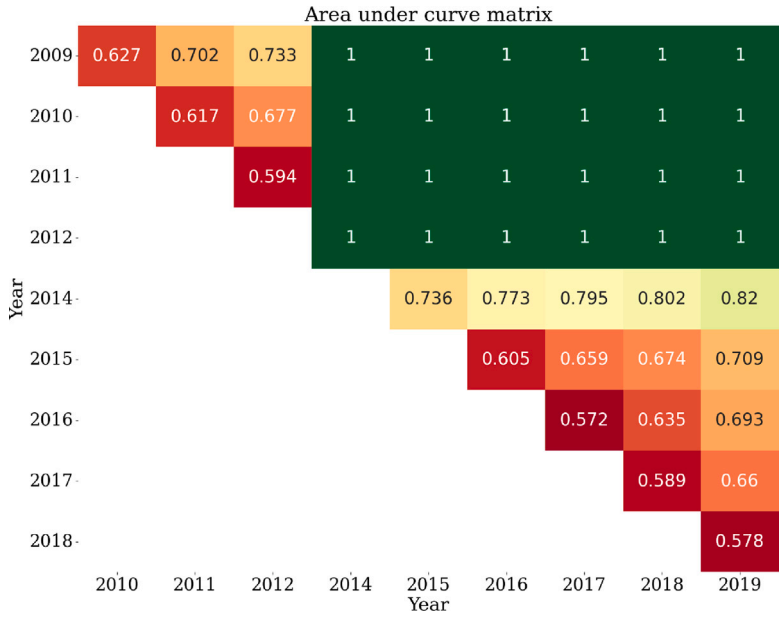

Fig. 4. Area under the curve matrix of DistilBERT text classification models predicting the year from which the data originated. An abrupt covariate shift is observed between the 2009–2012 data batch and the 2014–2019 data batch.

This drop is less severe in the 2009–2012 models but is still noticeable. Thus, the existence of concept shifts is confirmed.

Additionally, it can be observed that training and evaluating the model within the same year—as shown in the diagonal of the F1-score performance matrix—generally results in better performance compared to cross-year evaluations, although with some notable exceptions. The most pronounced exception is observed for the year 2014, where the model exhibits its poorest performance in the year it was trained. Similarly, there are instances, such as in 2009 and 2017, where the model's performance is marginally better in years other than the training year.

4.2. Continual learning

4.2.1. Backward transfer

The backward transfer, measured by the F1-score, for the CL strategies and baseline techniques is presented in Fig. 6. The x-axis represents the model's performance in a specific year, while the y-axis indicates the average F1-score obtained when testing the model with data from previous years.

As depicted in Fig. 6, CL strategies prevent significant performance drops compared to not utilizing CL techniques. All CL strategies perform above the expected lower performance bound defined by single fine-tuning. Furthermore, there is a clear trend of decreasing average F1-scores over time, with a more pronounced drop in 2015 when 2014 is included in the backward transfer computation. However, the performance decrease is moderate rather than severe.

When comparing the different techniques, joint training stands out as the approach offering the best overall performance over time in terms of backward transfer measured by the F1-score. It serves as the upper baseline, as expected. The static baseline exhibits a performance decrease over time, although not as severe as some of the other approaches assessed. Among the CL strategies, the cumulative approach performs the best, while single fine-tuning represents the lower performance bound in terms of backward transfer.

Table 5 presents the global backward transfer, computed for the AUC, accuracy, recall, precision, and F1-score, along with their 95% confidence intervals.

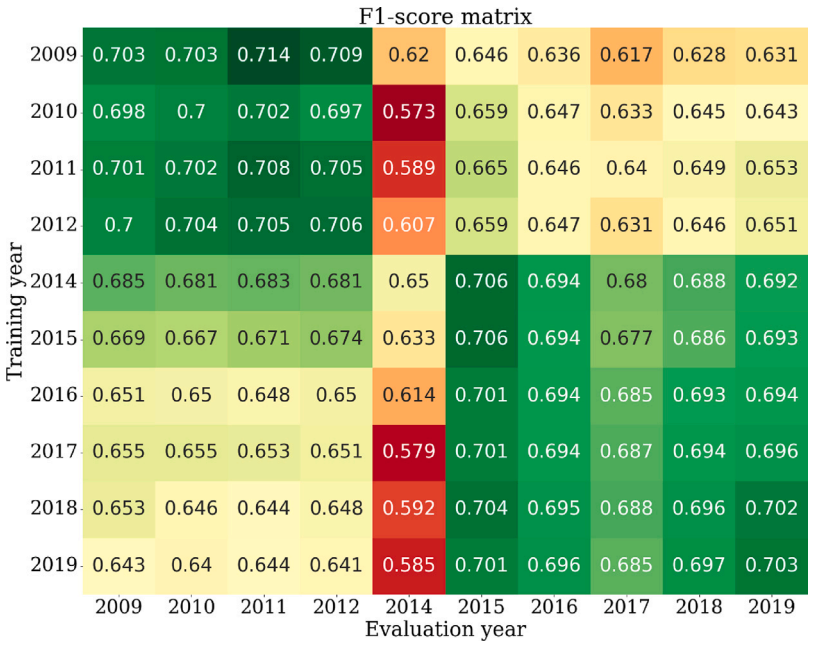

Fig. 5. F1-score matrix of DistilBERT text classification models trained on data from one year (y-axis) and evaluated on the test set of all years (x-axis). A moderate performance drop is observed between the 2009–2012 batch and the 2015–2019 batch, with the year 2014 showing the lowest performance.

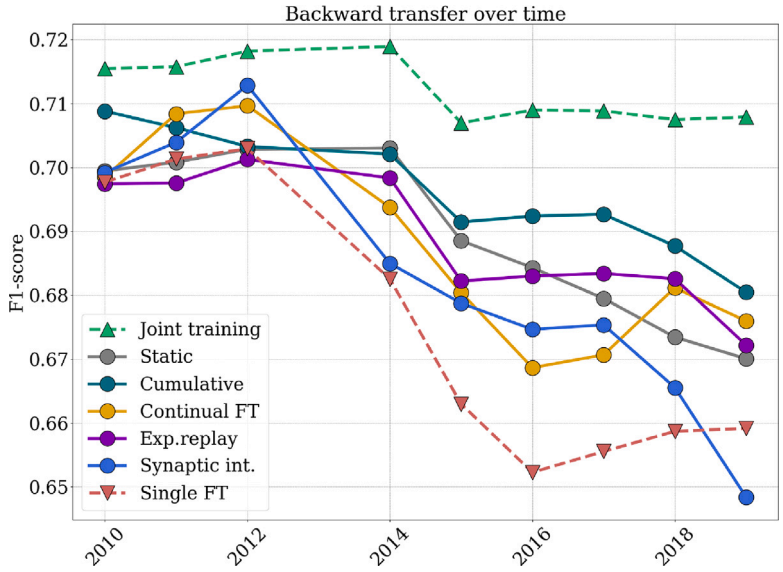

Fig. 6. Backward transfer over time, spanning from 2010 to 2019 (excluding 2009 due to the lack of available data for 2008) computed using the F1-score. CL strategies enhance knowledge retention over time. Abbreviations: FT, fine-tuning; Exp, experience; int, intelligence.

Table 5
Global backward transfer for each reference metric, with 95% confidence intervals shown in brackets. Abbreviations: AUC, area under curve; JT, joint training; ST, static; CM, cumulative; CFT, continual fine-tuning; ER, experience replay; SI, synaptic intelligence; SFT, single fine-tuning.

Strategy	AUC	Accuracy	Recall	Precision	F1-score
JT	0.809[0.808,0.809]	0.766[0.764,0.767]	0.744[0.742,0.746]	0.685[0.684,0.686]	0.712[0.711,0.713]
ST	0.823[0.821,0.824]	0.737[0.733,0.74]	0.747[0.743,0.751]	0.647[0.64,0.654]	0.689[0.686,0.692]
CM	0.793[0.787,0.798]	0.755[0.754,0.757]	0.72[0.713,0.726]	0.677[0.675,0.68]	0.696[0.694,0.698]
CFT	0.8[0.795,0.805]	0.763[0.762,0.764]	0.672[0.663,0.681]	0.711[0.708,0.715]	0.687[0.684,0.691]
ER	0.753[0.747,0.758]	0.754[0.753,0.756]	0.698[0.692,0.703]	0.683[0.681,0.685]	0.689[0.686,0.691]
SI	0.809[0.805,0.813]	0.763[0.762,0.765]	0.657[0.645,0.67]	0.721[0.715,0.727]	0.683[0.678,0.687]
SFT	0.753[0.759,0.765]	0.76[0.759,0.761]	0.643[0.632,0.655]	0.72[0.715,0.724]	0.675[0.67,0.679]

Table 5 demonstrates statistically significant differences $\alpha = 0.05$ between the implemented CL pipelines and the lower baseline—since the 95% confidence intervals are not overlapping [48]—indicating that CL techniques lead to performance improvements compared to not utilizing them. Among the CL strategies, the cumulative approach exhibits the best overall performance.

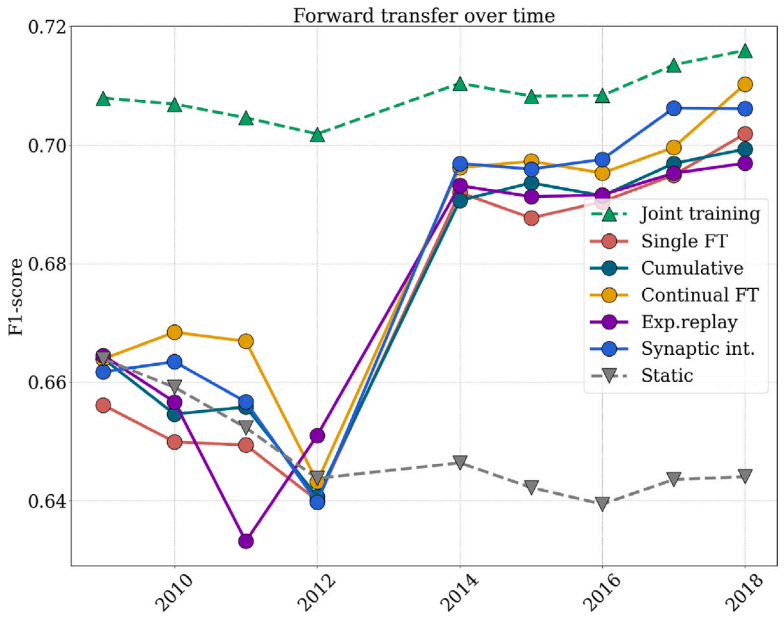

Fig. 7. Forward transfer over time, spanning from 2009 to 2018 (excluding 2019 due to the lack of available data for 2020) computed using the F1-score. CL strategies are crucial for enabling forward knowledge transfer over time. Abbreviations: FT, fine-tuning; Exp, experience; int, intelligence.

Table 6
Global forward transfer for each reference metric, with 95% confidence intervals shown in brackets. Abbreviations: AUC, area under curve; JT, joint training; SFT, single fine-tuning; CM, cumulative; CFT, continual fine-tuning; ER, experience replay; SI, synaptic intelligence; ST, static.

Strategy	AUC	Accuracy	Recall	Precision	F1-score
JT	0.818[0.817,0.819]	0.792[0.79,0.794]	0.724[0.723,0.725]	0.697[0.695,0.699]	0.709[0.708,0.71]
SFT	0.78[0.778,0.783]	0.755[0.746,0.765]	0.718[0.708,0.729]	0.644[0.631,0.657]	0.674[0.668,0.679]
CM	0.811[0.808,0.814]	0.749[0.739,0.758]	0.745[0.735,0.755]	0.628[0.616,0.641]	0.676[0.671,0.681]
CFT	0.814[0.811,0.817]	0.757[0.749,0.765]	0.743[0.737,0.75]	0.638[0.627,0.649]	0.682[0.677,0.687]
RP	0.789[0.784,0.795]	0.752[0.744,0.761]	0.732[0.721,0.743]	0.635[0.623,0.646]	0.675[0.67,0.68]
SI	0.823[0.819,0.826]	0.759[0.75,0.768]	0.73[0.722,0.737]	0.646[0.633,0.658]	0.68[0.675,0.686]
ST	0.811[0.811,0.812]	0.694[0.693,0.696]	0.808[0.805,0.811]	0.545[0.541,0.55]	0.648[0.646,0.65]

4.2.2. Forward transfer

Fig. 7 illustrates the forward transfer, measured by the F1-score, for the CL strategies and baseline techniques. The x-axis represents a specific year, and the y-axis indicates the average F1-score obtained when testing the model with data from the incoming years.

As observed in Fig. 7, CL strategies exhibit a distinct behavior compared to the baselines. The CL techniques show a common trend, with a notable increase in forward transfer in 2014. On the other hand, the baselines demonstrate the expected upper and lower bounds, with joint training serving as the upper bound and the static approach as the lower bound. Among the CL strategies, there is no clear winner as they interconnect over time, although continual fine-tuning and synaptic intelligence appear to perform well.

Table 6 presents the global forward transfer, computed for the AUC, accuracy, recall, precision, and F1-score, along with their 95% confidence intervals.

Table 6 indicates statistically significant differences ($\alpha = 0.05$) between the implemented CL pipelines and the lower baseline—since the 95% confidence intervals are not overlapping [48]—implying that CL techniques lead to improvements compared to not utilizing them. Among the CL strategies, continual fine-tuning and synaptic intelligence stand out.

4.3. Impact on the health service department of the Valencian Region

In Fig. 8, the accuracy per patient ICD-9 is illustrated over time, comparing the in-house triage protocol of the Valencian Region with the text model trained using the continual fine-tuning strategy, which provides the optimal balance between performance and efficiency. Our analysis concentrates on the top 5 most common and severe ICD-9 codes, as well as the top 5 most common but non-severe ICD-9 codes, offering a comprehensive perspective on the potential contributions in both case types.

Upon analyzing Fig. 8, the substantial value provided by our deep continual approach becomes evident. In the majority of incident types, the model's recommendations outperform those of the triage tree. Particularly noteworthy is the significant difference observed in the most severe cases, where the model clearly demonstrates superior performance compared to the in-house triage protocol. On the other hand, the protocol tends to perform slightly better for certain types of non-severe incidents. Hence, this figure underscores that under-triage is a concern for severe incidents and emphasizes the potential for our deep continual approach to significantly enhance patient care if implemented in the service.

Additionally, this figure enables the analysis of the evolution over time of the deep continual model alongside the in-house triage protocol. Concerning the model, it is noteworthy that the transition from 2012 to 2014 implies a slight performance decrease for severe incidents, while the performance of some non-severe incidents notably increased. Conversely, focusing on the in-house protocol, the transition from 2012 to 2014 tends to improve the assessment of the most severe incidents, while the performance regarding non-severe incidents worsened.

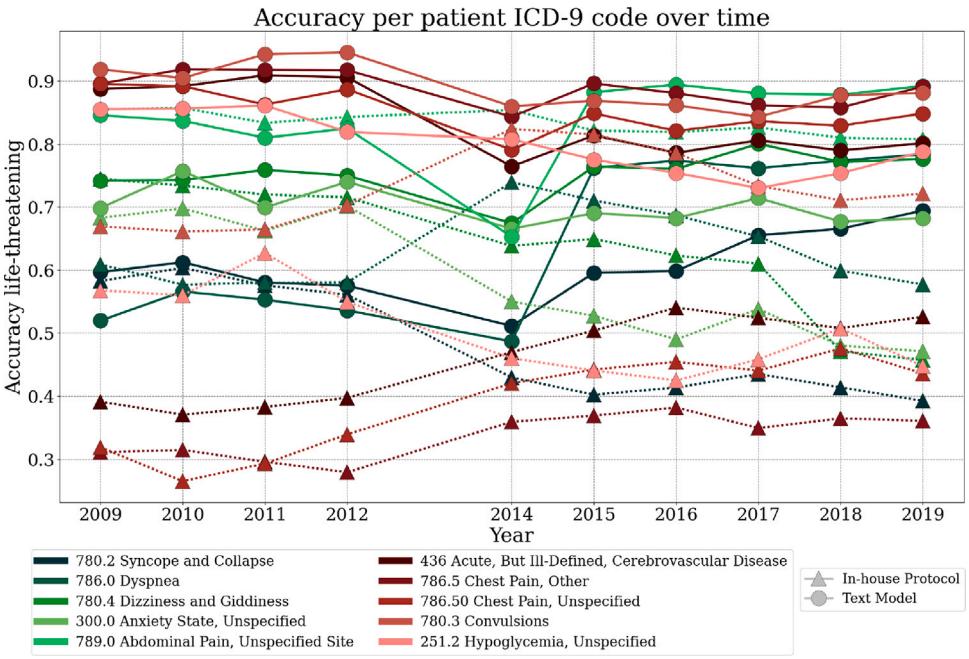

Fig. 8. Accuracy over time per patient ICD-9, comparing the in-house protocol with the text model employing the continual fine-tuning strategy. The analysis focuses on the five most frequent non-severe ICD-9 codes (green color scale) and the five most frequent severe ones (red color scale).

5. Discussion

5.1. Relevance

The findings of our study underscore the criticality of employing CL strategies for effective backward and forward knowledge transfer. To ensure the sustained performance of our EMCI classifier over time, the utilization of CL techniques becomes imperative. Importantly, our study represents the first investigation to incorporate CL within the learning pipelines of deep models designed for emergency triage support.

The identified dataset shifts, encompassing prior probability shifts, covariate shifts, and concept shifts, align closely with the changes implemented by the Health Services Department of the Valencian Community in 2014 regarding information system and coordination protocols. While these shifts may not be drastic, they are notable and should not be disregarded. Particularly, significant shifts in data distribution, pertaining to concepts and application-based data, can severely impede model performance. Consequently, our argument follows that the capacity to effectively handle these moderate yet significant data shifts may enhance model resilience when faced with more substantial changes in the future.

Fig. 9 provides a comprehensive overview of the global backward and forward metrics, specifically focusing on F1-score, as discussed in the previous section. This visual representation clearly demonstrates the indispensability of CL strategies in mitigating catastrophic forgetting, as they facilitate the accumulation of knowledge over time, while also enabling effective knowledge forward transfer.

Among the different continual learning strategies evaluated, both the cumulative and experience replay approaches exhibit similar behavior. The cumulative strategy can be viewed as an experience replay technique with unlimited memory. These two strategies outperform the others in terms of knowledge retention. However, in terms of knowledge transfer, the synaptic intelligence and continual fine-tuning approaches showcase superior performance, as they yield more positive impacts on predictive performance in subsequent years.

Considering the specific nature of our problem, where forward transfer holds greater significance than backward transfer, and taking into account the computational resources required for training time and memory, it would be reasonable to lean towards adopting a continual fine-tuning approach to address our problem. This choice offers several advantages, including easier integration into the retraining routine associated with the model, which can be seamlessly embedded into a deployed decision support system for emergency triage.

The continual fine-tuning approach allows us to capitalize on its superior ability to facilitate knowledge transfer and enhance predictive performance in subsequent years. Additionally, it offers practical benefits in terms of computational efficiency and resource utilization, which are valuable considerations when dealing with the constraints of our problem. By selecting this approach, we can effectively balance the importance of forward transfer, the available computational resources, and the ease of integration into our existing model retraining processes.

With a focus on the health-related implications of our work, it has been observed that the in-house triage protocol exhibits under-triage issues. This might stem from its structured nature, which implies a significantly more limited representation space compared to the flexibility offered by free text. The high uncertainty and tight time constraints associated with structured information gathering contribute to these errors. We posit that the flexibility and rapidity of free text data gathering make it a superior representation of incidents. When coupled with the temporal adaptation capabilities provided by a CL approach, our deep continual model emerges as an invaluable tool in the emergency medical triage context.

Specifically, the deep continual text model demonstrates significant potential in mitigating under-triage by recommending a review of under-triaged severe incidents, all while maintaining commendable performance in non-severe cases. Regarding these non-severe cases, the triage protocol slightly outperforms for some of them. However, since the real severity of an incident is unknown when the caller interacts with the dispatcher, considering our deep continual text model's recommendations constitutes a prudent approach, particularly given its notable value in severe incidents. Moreover, it is crucial to emphasize that in future developments, this deep continual text model will be complemented with other models integrating additional modalities, anticipating an enhancement in its overall performance. Consequently, our developments contribute to improvements in the value added to the patient and the health system.

In comparing the deep continual text models developed in this study with the previous DeepEMC^2 model, a direct assessment is not

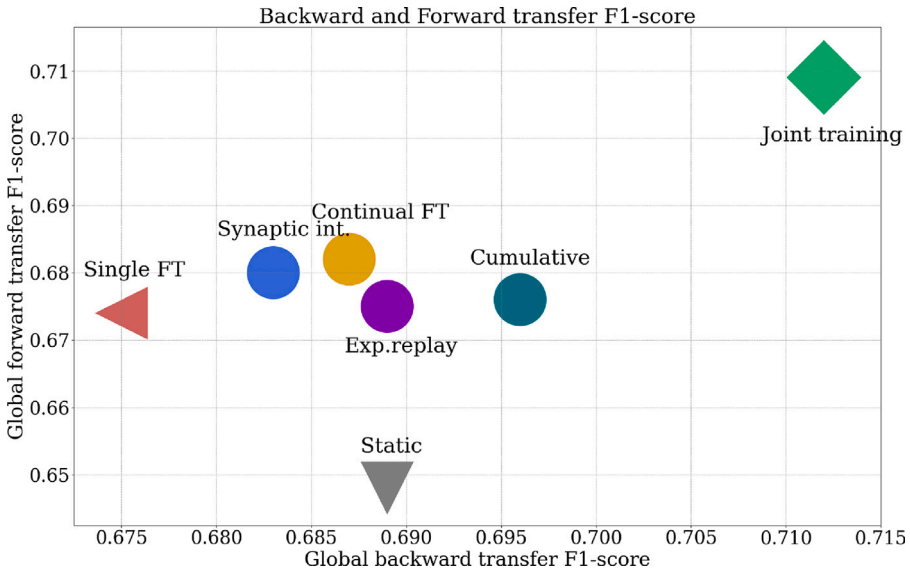

Fig. 9. Global backward and forward transfer, computed with the F1-score. CL strategies play a vital role to enhance backward and forward knowledge transfer. Abbreviations: FT, fine-tuning; Exp, experience; int, intelligence.

straightforward. This is due to several factors: DeepEMC^2 was not trained using a CL approach, its training data was exclusively from the 2009–2012 data batch, and it is both a multimodal and multitask model. However, focusing on Fig. 5, it is observed that for the years 2009 to 2012, the F1-score for the life-threatening label range between 0.69 and 0.7. Considering that the overall F1-score of DeepEMC^2 for the same period and label is 0.705 [17], this comparison underscores the significant contribution of free text dispatcher observations to the prediction. Furthermore, it demonstrates that utilizing DistilBERT models, which have fewer parameters than DeepEMC^2, does not result in a substantial loss in performance for the life-threatening label over the 2009–2012 period.

Finally, the observed performance fluctuations from transitioning from 2012 to 2014 for both the in-house triage protocol and the deep text model could be attributed to the introduction of novel dispatchers and changes within the protocol in the year 2013. These changes resulted in a more conservative triage approach for the protocol, reducing under-triage but increasing over-triage compared to the previous time period. Additionally, the slight increase in under-triage coupled with a decrease in over-triage observed in the text model from 2012 to 2014 could be associated with the quality and expressiveness of the registered data.

5.2. Limitations

The primary limitation of our work lies in the significant uncertainty associated with the phone triage process. Since data collection occurs remotely, within a time-critical context, the information gathered is often incomplete. Consequently, any model involved in providing decision support must rely on limited incoming data, which can introduce biases in certain cases. This inherent challenge imposes constraints on the achievable performance of any machine learning support model.

Another significant practical difficulty included the short length of the text fields in many cases. Consequently, the model had to cope with limited data. Additionally, there was a considerable computational cost associated with training a deep model across multiple learning experiences, under various configurations related to CL strategies, and with different combinations of hyperparameters.

5.3. Future work

Regarding future endeavors, we identify multiple directions for further exploration. Firstly, we propose a multitask CL approach to address the problem, incorporating considerations for admissible response delays and the jurisdiction labels of the emergency system. Secondly, we suggest the inclusion of additional input features, such as demographics, contextual information, or structured clinical features, thus forming a multimodal CL approach. These avenues of research hold promise for advancing the field and expanding the scope of our investigations. Finally, the deep continual models developed in this work could serve as a starting point for developing models dealing with similar datasets from other emergency medical dispatch centers.

6. Conclusions

In this work, we have conducted an extensive investigation into dataset shifts and CL strategies within the domain of EMCI triage. Our study provides compelling evidence of prior probability shifts, covariate shifts, and concept shifts within our data, which directly impact the performance of models over time. The utilization of CL strategies has been demonstrated as crucial in mitigating the adverse effects caused by distributional drifts, both in terms of backward and forward knowledge transfer. Consequently, embracing a CL approach proves highly valuable as it ensures the retention of pertinent knowledge while demonstrating adaptability to novel patterns. These attributes result in improved model performance, thereby enhancing the quality of the potential decision-support process facilitated by such models, particularly within the context of EMCI triage. Considering the advanced capabilities of deep continual text models in assessing the life-threatening level of the most severe incidents, demonstrating significantly lower under-triage rates in these cases, the implementation of the CL routines developed in this study for EMCI triage will have a significant and positive direct impact on patient well-being and the sustainability of health services.

CRediT authorship contribution statement

Pablo Ferri: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Vincenzo Lomonaco:** Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Lucia C. Passaro:** Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Antonio Félix-De Castro:** Validation, Supervision, Resources, Project administration,

Data curation, Conceptualization. **Purificación Sánchez-Cuesta:** Supervision, Resources, Project administration, Data curation, Conceptualization. **Carlos Sáez:** Supervision, Project administration, Investigation, Formal analysis. **Juan M. García-Gómez:** Supervision, Project administration, Investigation, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has received support from the Ministry of Science, Innovation, and Universities of Spain through the FPU18/06441 program. In addition, it has been partly funded by PNRR-M4C2-Investimento 1.3, Partenariato Esteso PE00000013-FAIR-Future Artificial Intelligence Research-Spoke 1 Human-centered AI, funded by the European Commission under the NextGeneration EU programme.

References

[1] J.J. Clawson, K.B. Dernocoeur, Principles of emergency medical dispatch, 1988.
[2] G. FitzGerald, G. Jelinek, D. Scott, M. Gerdtz, Emergency department triage revisited, Emerg. Med. J. 27 (2) (2010) 86–92.
[3] M. Storm-Versloot, D. Ubbink, J. Kappelhof, J. Luitse, Comparison of an informally structured triage system, the emergency severity index, and the manchester triage system to distinguish patient priority in the emergency department, Acad. Emerg. Med. 18 (8) (2011) 822–829.
[4] R. Wuerz, D. Travers, N. Gilboy, D. Eitel, A. Rosenau, R. Yazhari, Implementation and refinement of the emergency severity index, Acad. Emerg. Med.: Off. J. Soc. Acad. Emerg. Med. 8 (2) (2001) 170–176.
[5] K. Mackway-Jones, J. Marsden, J. Windle, Emergency Triage: Manchester Triage Group, John Wiley & Sons, 2013.
[6] J. Barroeta Urquiza, N. Boada Bravo, Los servicios de emergencia y urgencias médicas extrahospitalarias en España, Mensor (2011).
[7] L. Tollinton, A.M. Metcalf, S. Velupillai, Enhancing predictions of patient conveyance using emergency call handler free text notes for unconscious and fainting incidents reported to the London ambulance service, Int. J. Med. Inform. 141 (2020) 104179.
[8] P. Ferri, C. Sáez, A. Félix-De Castro, P. Sánchez-Cuesta, J. García-Gómez, Discovering key topics in emergency medical dispatch from free text dispatcher observations, Stud. Health Technol. Inform. 294 (2022) 859–863.
[9] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[10] Z.-H. Zhan, J.-Y. Li, J. Zhang, Evolutionary deep learning: A survey, Neurocomputing 483 (2022) 42–58.
[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (2015) 211–252.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[13] G.H. De Rosa, J.P. Papa, A survey on text generation using generative adversarial networks, Pattern Recognit. 119 (2021) 108098.
[14] D. Spangler, T. Hermansson, D. Smekal, H. Blomberg, A validation of machine learning-based risk scores in the prehospital setting, PLoS One 14 (12) (2019) e0226518.
[15] S.N. Blomberg, F. Folke, A.K. Ersbøll, H.C. Christensen, C. Torp-Pedersen, M.R. Sayre, C.R. Counts, F.K. Lippert, Machine learning as a supportive tool to recognize cardiac arrest in emergency calls, Resuscitation 138 (2019) 322–329.
[16] R. Inokuchi, M. Iwagami, Y. Sun, A. Sakamoto, N. Tamiya, Machine learning models predicting undertriage in telephone triage, Ann. Med. 54 (1) (2022) 2989–2996.
[17] P. Ferri, C. Sáez, A. Félix-De Castro, J. Juan-Albarracín, V. Blanes-Selva, P. Sánchez-Cuesta, J. García-Gómez, Deep ensemble multitask classification of emergency medical call incidents combining multimodal data improves emergency medical dispatch, Artif. Intell. Med. 117 (2021) 102088.
[18] J. Quinonero-Candela, M. Sugiyama, A. Schwaighofer, N. Lawrence, Dataset Shift in Machine Learning, MIT Press, 2008.
[19] J. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. Chawla, F. Herrera, A unifying view on dataset shift in classification, Pattern Recognit. 45 (1) (2012) 521–530.
[20] M. McCloskey, N. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, in: G. Bower (Ed.), Psychology of Learning and Motivation, Vol. 24, Academic Press, 1989, pp. 109–165.
[21] G. Parisi, R. Kemker, J. Part, C. Kanan, S. Wermter, Continual lifelong learning with neural networks: A review, Neural Netw. 113 (2019) 54–71.
[22] V. Lomonaco, L. Pellegrini, A. Cossu, A. Carta, G. Graffieti, T. Hayes, M. Lange, M. Masana, J. Pomponi, G. Ven, M. Mundt, Q. She, K. Cooper, J. Forest, E. Belouadah, S. Calderara, G. Parisi, F. Cuzzolin, A. Tolias, D. Maltoni, Avalanche: An end-to-end library for continual learning, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, CVPRW, 2021, pp. 3595–3605.
[23] L.L. Guo, S.R. Pfohl, J. Fries, A.E. Johnson, J. Posada, C. Aftandilian, N. Shah, L. Sung, Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine, Sci. Rep. 12 (1) (2022) 2726.
[24] J. Lemmon, L.L. Guo, J. Posada, S.R. Pfohl, J. Fries, S.L. Fleming, C. Aftandilian, N. Shah, L. Sung, Evaluation of feature selection methods for preserving machine learning performance in the presence of temporal dataset shift in clinical medicine, Methods Inf. Med. 62 (01/02) (2023) 060–070.
[25] L.L. Guo, E. Steinberg, S.L. Fleming, J. Posada, J. Lemmon, S.R. Pfohl, N. Shah, J. Fries, L. Sung, EHR foundation models improve robustness in the presence of temporal distribution shift, Sci. Rep. 13 (1) (2023) 3767.
[26] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (4) (2020) 1234–1240.
[27] S. Lee, C. Yin, P. Zhang, Stable clinical risk prediction against distribution shift in electronic health records, Patterns 4 (9) (2023).
[28] G. Rossum, Python reference manual, in: Department of Computer Science CS, 1995, Issue R 9525. CWI.
[29] S. Walt, S. Colbert, G. Varoquaux, The numpy array: A structure for efficient numerical computation, Comput. Sci. Eng. 13 (2) (2011) 22–30.
[30] W. McKinney, Data structures for statistical computing in python, 2010, pp. 56–61.
[31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch., 2017.
[32] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Scao, S. Gugger, A. Rush, Huggingface's transformers: state-of-the-art natural language processing, 2020, (arXiv:1910.03771). arXiv.
[33] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2623–2631.
[34] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al., Google's neural machine translation system: Bridging the gap between human and machine translation, 2016, arXiv preprint arXiv:1609.08144.
[35] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT, 2020, arXiv:1910.01108. arXiv.
[36] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019, arXiv:1810.04805. arXiv.
[37] D. Kwiatkowski, P. Phillips, P. Schmidt, Y. Shin, Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? J. Econometrics 54 (1) (1992) 159–178.
[38] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, 2015, arXiv:1503.02531 [Cs, Stat]..
[39] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, 2023, arXiv preprint arXiv:2303.08774.
[40] J. Ba, J. Kiros, G. Hinton, Layer normalization, 2016, Cs,Stat.
[41] G. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, 2012, arXiv:1207.0580. arXiv.
[42] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019, (arXiv:1711.05101). arXiv.
[43] D. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017, arXiv:1412.6980 [Cs]..
[44] K. Janocha, W. Czarnecki, On loss functions for deep neural networks in classification, 2017, arXiv:1702.05659. arXiv.
[45] D. Lopez-Paz, M. Ranzato, Gradient episodic memory for continual learning, 2017.
[46] F. Zenke, B. Poole, S. Ganguli, Continual learning through synaptic intelligence, 2017, arXiv:1703.04200 [Cs, q-Bio, Stat].
[47] B. Settles, Active learning literature survey, 2009.
[48] B. Rosner, Fundamentals of Biostatistics, Cengage learning, 2015.