Continual learning of longitudinal health records

Jacob Armstrong
Institute of Biomedical Engineering
Oxford University
jacob.armstrong@eng.ox.ac.uk

David A. Clifton
Institute of Biomedical Engineering
Oxford University
davidc@robots.ox.ac.uk

2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI) | 978-1-6654-8791-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/BHI56158.2022.9926878

***Abstract*—*Continual learning* denotes machine learning methods which can adapt to new environments while retaining and reusing knowledge gained from past experiences. Such methods address two issues encountered by models in non-stationary environments: ungeneralisability to new data, and the catastrophic forgetting of previous knowledge during retraining. This is a pervasive problem in clinical settings where patient data exhibits covariate shift not only between populations, but also continuously over time. However, while continual learning methods have seen nascent success in the imaging domain, they have been little applied to the multi-variate sequential data characteristic of patient recordings. Here we evaluate a variety of continual learning methods on longitudinal ICU data in a series of representative healthcare scenarios. We find that while several methods mitigate short-term forgetting, domain shift remains a challenging problem over a large series of tasks, with only replay based methods achieving stable long-term performance.**

***Index Terms*—Continual learning, domain adaptation, time series, clinical machine learning, EHR**

I. Introduction

Clinical and healthcare-related machine learning studies have grown rapidly in recent years, with over a thousand publications annually since 2018 [1]. However many models suffer from ungeneralisability: the distribution of their training data is not representative of the setting in which they are deployed, and hence their real-world performance and utility is overestimated. Further, the distribution of data in a given environment itself continually shifts with time, limiting the use even of models trained on initially representative domains [2, 3].

Unfortunately, naively retraining networks on new data as it becomes available ("fine tuning") commonly results in forgetting of past knowledge. Models can overfit to the specific features of the new dataset, degrading performance on previous tasks in a process known as *catastrophic forgetting*. This occurs since training on the current task propels updated parameter values far from the previously optimized values (see Fig 1). This effectively overwrites learned features pertinent to previous tasks when they are not useful for the current one. While accumulating data and periodically retraining models theoretically alleviates catastrophic forgetting, such approaches are practically encumbered by privacy, storage, and computational hurdles.

Continual learning (CL) has recently emerged as a field to tackle these issues. Models are designed to incrementally update on new datasets while retaining and reusing past knowledge where relevant. Concretely this refers to models which can sequentially train on a series of tasks, while retaining predictive power on previously encountered examples.

However a number of state of the art techniques rely on storing past examples and hence may be infeasible in clinical settings due to privacy or data storage limitations. Generative models which create simulated pseudo-examples face further issues of computational limitations. Further, while a large proportion of Electronic Health Records (EHR) consist of periodic tabular readings (i.e. multi-variate time-series), most evaluations of continual learning methods are in the image domain [4, 5]. Current benchmarks do not adequately capture the realistic issues faced in a clinical context (e.g. highly imbalanced classes, large multivariate sequences, sparse recordings) [6], and hence the generalisability of their results to these contexts is unclear.

Contributions: In this work we present a set of representative continual learning scenarios in the medical domain derived from the open-access eICU-CRD and MIMIC-III ICU datasets [7, 8, 9]. We evaluate a range of methods on these problems, the first (to our knowledge) comprehensive study of Continual Learning methods on medical time-series data. Benchmarks demonstrate common domain shifts encountered by clinical systems in the real world, across geographies, time, and population demographics.

Related work: Cossu et al. [10] present a comprehensive evaluation of methods on a set of proposed benchmarks for sequence data. We extend on this work by evaluating such methods on real-world clinical scenarios, over a broader array of model architectures. Aljundi et al. [11] examine imbalanced classification problems and the effect of dropout regularisation but from a task incremental perspective on imaging data. Kiyasseh et al. [12] investigate domain incremental learning on univariate physiological signals but examine only replay based methods. Churamani et al. [13] investigate domain incremental learning across ethnicity and gender but for facial image data, only evaluating regularization based methods. Guo et al. [14] and Alves et al. [15] investigate temporal and institutional domain shift in ICU data, but from a domain adaptation perspective, considering only a single source and target dataset.

II. Background

A. Continual Learning Scenarios

The typical continual learning problem consists of a model encountering a sequence of discrete batches of data, corresponding to different ‘tasks’, where data cannot be stored between tasks. For example a clinical decision model updated annually on new hospital data. The data cannot be retained longer than this due to privacy limitations, but we aspire for the model to generalise to the population with each dataset encountered, and not overfit to the most recent batch as is seen in traditional supervised learning.

We limit our experiments to ‘Domain Incremental’ continual learning problems [16], where the task is nominally the same but the distribution of input-features changes with each task.

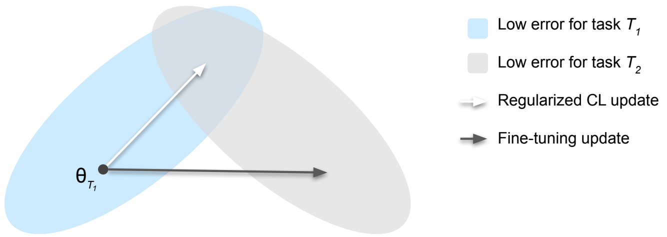

Fig. 1: Under naive transfer learning (grey arrow), there is no guarantee that the parameter values (θ_{T_1}) remain within a region of low error for the previous task T_1 (blue oval) after training on subsequent task T_2. Regularization techniques like Elastic Weight Consolidation (EWC) enforce such behaviour by penalising the loss, constraining parameter updates to a locus of learned values for previous tasks (figure adapted from [17]).

B. Ontology of methods

A number of methods have been proposed in recent years to mitigate catastrophic forgetting, falling under three general archetypes [18]:

Regularization: A regularization constraint is added to the loss function, enforcing updated parameter values to lie within a radius of the current value. This has the benefit of a natural Bayesian interpretation where the posterior values after training on task T_i inform the priors for task T_{i+1}. Methods differ in strategies for choosing which parameters to constrain, and to what degree.

Rehearsal: A subset of examples (or generated pseudo-examples) from previous tasks are cached and mixed in with each new task’s training set. Methods differ chiefly in the criteria used for choosing examples. Also known as *replay*.

Dynamic architectures: A broad variety of techniques where the network architecture itself adapts with new task presentation. Approaches range from task conditioned hyper-networks, to networks which add neurons as resources are required to model new tasks.

Such architectural features are not mutually exclusive. For example, GEM [19], iCARL [20], and FRoMP [21] employ both rehearsal and regularisation elements. More complex ontologies have been proposed for finer categorisation [6].

Rehearsal methods achieve state of the art in many scenarios examined in the literature [22, 23]. However, such techniques are often infeasible in real-world settings, where previous examples cannot be stored or shared due to data privacy constraints [4, 24]. Such a problem is not unique to clinical settings, and while *generative* replay models simulating past examples have been proposed [25], sparse and complex sequential data can prohibit learning of an adequate generative distribution function [26].

For an in depth review of continual learning methods generally, we refer to Delange et al. [6], Parisi et al. [18], Luo et al. [27]. For convenience, we briefly outline the methods evaluated in this work below:

Regularization approaches:

- **Elastic Weight Consolidation (EWC)** [17] Penalises changes in parameter values relative to the *importance* of parameters to previous task(s). Importance measured via Fisher’s information matrix.
- **Online EWC** [28] An adaptation of EWC using a running average of task importance penalties, as opposed to distinct penalties for each previous task. Computationally more efficient and tractable for a large number of tasks.
- **Synaptic Intelligence (SI)** [29] Similar to EWC, enforces parameter specific regularization but importances are calculated *online* (i.e. during training) by approximating the effect on loss and gradient update, as opposed to during an additional pass of the network post training.
- **Learning without Forgetting (LwF)** [30] A copy of the model parameters before updating on the current task is stored and compared to the updated version. Parameter values are distillepd between both versions for final update. Hence may be categorised as *functional* regularization.

Replay approaches:

- **Replay** Naive storage of a set of random examples per task, which are mixed in with each subsequent task’s training data. May employ more specific storage policies such as class or task-wise balancing of memories.
- **Gradient Episodic Memory (GEM)** [19] Stores a set of examples from each task. Selectively updates gradient for a given minibatch on the current task only if the gradient can be projected in a plane which maintains the positivity of the gradient updates for all stored examples.
- **Averaged GEM (A-GEM)** [31] Adaptation of GEM considering only the average gradient for a randomly sampled subset of the stored examples.

III. Experiments

A. Problem definitions

We consider 3 natural Domain Incremental experiments, corresponding to n patient ICU datasets encountered sequentially across time or location. Domain shifts correspond to changing: season $(n = 4)$; region $(n = 4)$; and hospital $(n = 155)$.

We also consider 3 artificial Domain Incremental experiments, simulating imbalanced populations between healthcare environments (due to demographic-specific care in a given institution, or general population imbalance). Domain increments correspond to groups of patients split by: age group $(n = 7)$; ethnicity $(n = 5)$; and ICU ward type $(n = 8)$.

For each task the setting is supervised prediction of a binary outcome (48hr in-hospital mortality). Input data are multivariate time-series, consisting of periodically recorded patient vital

Archetype	Method	Abbreviation	Source
Baseline	Naive fine-tuning	Naive	
	Cumulative multi-task training	Cumulative	
Regularization	Elastic Weight Consolidation	EWC	[17]
	Online EWC	Online EWC	[28]
	Synaptic Intelligence	SI	[29]
	Learning without Forgetting	LwF	[30]
Rehearsal	Naive replay	Replay	
	GDumb	GDumb	[33]
	Gradient Episodic Memory	GEM	[19]
	Averaged Gradient Episodic Memory	AGEM	[31]

TABLE I: Continual Learning methods evaluated.

MIMIC-III	eICU	Domain increment	Number of domains
	✓	Region (US)	4
	✓	Hospital	155
✓	✓	ICU Ward	5-8
✓	✓	Age	6-7
✓	✓	Ethnicity (broad)	5
✓		Season	4

TABLE II: Domain shifts annotated in the MIMIC-III and eICU-CRD datasets. Ranges of values correspond to different domain splits in each dataset.

signs from an ICU admission. These are sampled hourly, and are of duration $t = 48$ time steps. Static covariates are repeated to the length of the time-varying sequence and concatenated to enable processing by sequential models.

B. Experimental setup

Model architectures: For each problem, we evaluate 4 basic neural network architectures: a dense feedforward network (MLP); 1d convolutional neural network (CNN); long- short-term memory network (LSTM); and a transformer. These were chosen to give a breadth of sequential models, along with a data-structure agnostic model (MLP) for baseline comparison. Models consist of one to four architecture-specific layers, followed by two dense linear layers. Standard regularization features such as dropout were omitted to clearer investigate the effect of the continual learning mechanisms themselves. Batch Normalisation was not used in the CNN due to its intensifying effect on catastrophic forgetting [32].

Strategies: Each model is equipped with one of the 7 continual learning strategies listed in Table I. Rehearsal based methods are given a fixed budget of 256 samples per task, corresponding to approximately 5% and 0.5% of the training data for MIMIC and eICU experiments respectively.

We further evaluate all models against two baseline methods:

- **Naive:** Naive fine-tuning on each additional task. This is a soft lower bound on performance, equivalent to serial transfer learning with no continual learning mechanism.
- **Cumulative:** Cumulative multi-task retraining on all tasks seen thus far. This is a soft upper bound on performance.

Data: We use the open-access eICU-CRD [8] dataset for all experiments bar seasonal and narrow ethnicity domain increments, for which such information was not available. For these we use the open-access MIMIC-III [7, 9] ICU database. For standardisation of preprocessing and outcome definitions, datasets were preprocessed with the FIDDLE pipeline [34]. Data can be accessed at https://www.physionet.org/content/mimic-eicu-fiddle-feature/1.0.0/. Code for reproducing all experiments can be found at https://github.com/iacobo/continual.

Relevant domain shifts identifiable in both datasets are listed in Table II.

Metrics: Since class sizes are highly imbalanced in all experiments (mortality outcome averaging 10% across tasks), and the degree of class imbalance is not constant across domain splits, accuracy is an inappropriate measure of model performance [35]. In minority-event detection, metrics such as sensitivity and specificity are often preferred depending on the relative importance of Type I and Type II errors in the given medical context [36]. To simplify presentation of results, we report the Balanced Accuracy, an average of specificity and sensitivity.

Pipeline:

1) **Task split** Data is initially split into several tasks stratified by patient demographic. Task order was randomized.
2) **Train, validation, test split** Data within each task is then split into train (85%) and validation (15%) sets for the first two tasks, and train (85%) and test (15%) sets for subsequent tasks. Since multiple ICU admissions can pertain to the same patient, train/validation/test streams were split along patient identities to avoid data leakage of similar records [37].
3) **Hyperparameter optimisation** Hyperparameter optimisation requires careful consideration in a continual learning setting, since we should not have access to validation sets from future tasks during the model specification phase. As such, tuning was performed using validation data from the first two tasks only. This setup is consistent with validation regimes proposed in [31]. Generic hyperparameters (i.e. learning rate, batch size, number of layers, hidden depth) were tuned for the Naive baseline run only and frozen for all other methods. Strategy specific hyperparameters were tuned independently for each method. Hyperparameters were sampled from a range of reasonable values determined from the literature [34, 10]. Where methods shared identical or analogous parameters, the search-space was also shared to ensure fair comparison. Hyperparameters were chosen to maximise the average balanced accuracy of the validation predictions for the first two tasks.
4) **Training** Once hyper-parameters were selected, each model/strategy combination was trained from scratch on the sequence of tasks' training data. The objective function of training was minimising the weighted cross entropy of predictions. Weights are determined by the inverse proportion of class examples in the first two tasks' training data.
5) **Evaluation** Models were evaluated on balanced accuracy for each task's test data. Per-task and average metrics were recorded at the end of each training epoch. Training and evaluation was repeated from random initialisation 5 times. Mean performance and bootstrapped 95% confidence intervals are recorded.

		Age				Ethnicity (broad)				ICU Ward				Season			
		CNN	LSTM	MLP	Transformer	CNN	LSTM	MLP	Transformer	CNN	LSTM	MLP	Transformer	CNN	LSTM	MLP	Transformer
Baseline	Cumulative	64.9$_{\pm 1.6}$	64.3$_{\pm 0.8}$	63.0$_{\pm 6.4}$	59.0$_{\pm 4.9}$	61.6$_{\pm 1.1}$	60.6$_{\pm 1.2}$	60.6$_{\pm 0.8}$	53.7$_{\pm 4.8}$	59.2$_{\pm 0.9}$	58.2$_{\pm 1.8}$	58.8$_{\pm 1.0}$	57.3$_{\pm 4.1}$	64.1$_{\pm 1.1}$	65.9$_{\pm 1.4}$	65.5$_{\pm 2.0}$	53.5$_{\pm 6.8}$
	Naive	64.0$_{\pm 0.7}$	62.8$_{\pm 1.2}$	50.0$_{\pm 0.0}$	57.8$_{\pm 3.8}$	67.6$_{\pm 0.9}$	67.6$_{\pm 0.7}$	68.8$_{\pm 0.9}$	50.0$_{\pm 0.0}$	64.2$_{\pm 1.7}$	56.6$_{\pm 2.0}$	58.7$_{\pm 1.2}$	53.7$_{\pm 4.4}$	67.2$_{\pm 2.3}$	67.8$_{\pm 1.5}$	67.6$_{\pm 1.0}$	50.0$_{\pm 0.0}$
Regularization	EWC	63.8$_{\pm 1.3}$	63.6$_{\pm 0.9}$	50.0$_{\pm 0.0}$	55.7$_{\pm 5.7}$	67.3$_{\pm 1.1}$	66.8$_{\pm 1.6}$	69.1$_{\pm 0.4}$	53.5$_{\pm 6.8}$	62.9$_{\pm 1.1}$	58.8$_{\pm 3.5}$	58.6$_{\pm 1.3}$	51.9$_{\pm 3.7}$	66.4$_{\pm 1.2}$	66.9$_{\pm 0.9}$	68.0$_{\pm 0.6}$	50.0$_{\pm 0.0}$
	LwF	**64.7**$_{\pm 0.9}$	**64.4**$_{\pm 0.8}$	64.3$_{\pm 0.7}$	58.4$_{\pm 4.2}$	67.5$_{\pm 0.2}$	**67.8**$_{\pm 1.1}$	69.2$_{\pm 0.6}$	52.6$_{\pm 5.0}$	64.3$_{\pm 0.5}$	**61.8**$_{\pm 1.0}$	**62.4**$_{\pm 2.3}$	**54.5**$_{\pm 5.5}$	67.1$_{\pm 2.0}$	67.0$_{\pm 1.4}$	67.8$_{\pm 0.8}$	52.4$_{\pm 4.7}$
	OnlineEWC	63.7$_{\pm 0.8}$	63.1$_{\pm 0.6}$	**64.6**$_{\pm 0.5}$	**61.1**$_{\pm 0.7}$	**67.8**$_{\pm 0.5}$	66.7$_{\pm 1.7}$	**70.0**$_{\pm 0.8}$	53.2$_{\pm 6.3}$	64.2$_{\pm 0.8}$	59.5$_{\pm 2.4}$	58.9$_{\pm 1.9}$	50.0$_{\pm 0.0}$	67.7$_{\pm 1.1}$	**67.8**$_{\pm 0.5}$	68.1$_{\pm 0.4}$	50.0$_{\pm 0.0}$
	SI	63.9$_{\pm 1.4}$	62.8$_{\pm 1.9}$	63.7$_{\pm 0.3}$	57.3$_{\pm 4.3}$	67.5$_{\pm 1.0}$	67.3$_{\pm 1.6}$	69.9$_{\pm 0.4}$	54.8$_{\pm 6.5}$	**64.5**$_{\pm 0.5}$	58.9$_{\pm 1.1}$	60.6$_{\pm 1.8}$	50.0$_{\pm 0.0}$	66.1$_{\pm 0.6}$	67.6$_{\pm 0.6}$	67.6$_{\pm 0.6}$	52.7$_{\pm 5.4}$
Rehearsal	AGEM	64.5$_{\pm 1.0}$	62.2$_{\pm 0.9}$	64.1$_{\pm 0.6}$	58.0$_{\pm 4.1}$	64.8$_{\pm 2.2}$	67.3$_{\pm 1.5}$	68.7$_{\pm 0.2}$	**56.1**$_{\pm 7.3}$	63.9$_{\pm 1.3}$	59.2$_{\pm 1.5}$	60.8$_{\pm 0.9}$	53.9$_{\pm 4.7}$	**68.4**$_{\pm 1.6}$	67.2$_{\pm 2.1}$	**68.6**$_{\pm 0.9}$	50.0$_{\pm 0.0}$
	GEM	63.1$_{\pm 0.8}$	60.6$_{\pm 1.1}$	61.7$_{\pm 0.6}$	58.5$_{\pm 1.4}$	58.2$_{\pm 1.1}$	57.8$_{\pm 1.1}$	60.2$_{\pm 0.4}$	50.8$_{\pm 1.6}$	60.3$_{\pm 1.6}$	57.4$_{\pm 1.5}$	57.3$_{\pm 1.3}$	53.8$_{\pm 3.2}$	60.1$_{\pm 1.1}$	60.1$_{\pm 2.4}$	63.7$_{\pm 0.9}$	54.4$_{\pm 5.2}$
	Replay	60.0$_{\pm 1.2}$	58.1$_{\pm 1.8}$	51.1$_{\pm 2.2}$	59.0$_{\pm 1.6}$	61.6$_{\pm 3.7}$	60.3$_{\pm 3.6}$	61.6$_{\pm 2.1}$	51.5$_{\pm 3.0}$	59.0$_{\pm 1.7}$	55.7$_{\pm 1.6}$	58.7$_{\pm 1.5}$	53.2$_{\pm 3.8}$	65.9$_{\pm 3.0}$	61.4$_{\pm 2.3}$	65.2$_{\pm 1.8}$	**55.6**$_{\pm 4.7}$

		Hospital (7)			Hospital (14)			Hospital (21)			Hospital (28)			Hospital (35)		
		CNN	LSTM	MLP	CNN	LSTM	MLP	CNN	LSTM	MLP	CNN	LSTM	MLP	CNN	LSTM	MLP
Baseline	Cumulative	57.3$_{\pm 1.2}$	55.2$_{\pm 0.8}$	56.5$_{\pm 0.3}$	62.2$_{\pm 2.5}$	61.6$_{\pm 0.8}$	61.5$_{\pm 0.3}$	57.9$_{\pm 1.0}$	60.3$_{\pm 1.2}$	60.9$_{\pm 0.8}$	54.6$_{\pm 0.6}$	55.5$_{\pm 0.9}$	56.1$_{\pm 0.7}$	56.0$_{\pm 1.7}$	56.9$_{\pm 1.5}$	56.1$_{\pm 1.6}$
	Naive	52.6$_{\pm 0.1}$	52.4$_{\pm 0.3}$	55.0$_{\pm 0.1}$	57.4$_{\pm 1.4}$	57.9$_{\pm 1.8}$	61.9$_{\pm 0.8}$	58.3$_{\pm 1.8}$	57.0$_{\pm 0.9}$	61.1$_{\pm 0.9}$	52.0$_{\pm 0.5}$	52.6$_{\pm 0.7}$	54.1$_{\pm 0.4}$	52.2$_{\pm 0.4}$	52.0$_{\pm 0.4}$	52.5$_{\pm 0.1}$
Regularization	EWC	52.6$_{\pm 0.0}$	52.5$_{\pm 0.1}$	54.5$_{\pm 1.1}$	57.9$_{\pm 1.4}$	58.9$_{\pm 0.5}$	61.2$_{\pm 1.1}$	58.8$_{\pm 1.7}$	57.4$_{\pm 1.6}$	61.8$_{\pm 1.0}$	52.4$_{\pm 0.6}$	**54.7**$_{\pm 1.8}$	54.2$_{\pm 0.4}$	51.9$_{\pm 0.1}$	52.5$_{\pm 0.8}$	52.5$_{\pm 0.1}$
	LwF	52.6$_{\pm 0.1}$	52.6$_{\pm 0.1}$	55.0$_{\pm 0.1}$	56.6$_{\pm 0.4}$	57.4$_{\pm 1.0}$	61.1$_{\pm 1.0}$	58.8$_{\pm 1.2}$	57.8$_{\pm 1.0}$	61.8$_{\pm 0.9}$	51.8$_{\pm 0.5}$	53.9$_{\pm 0.9}$	54.1$_{\pm 0.6}$	51.9$_{\pm 0.1}$	51.8$_{\pm 0.3}$	52.4$_{\pm 0.0}$
	OnlineEWC	52.6$_{\pm 0.0}$	52.5$_{\pm 0.1}$	55.0$_{\pm 0.1}$	57.1$_{\pm 0.7}$	58.6$_{\pm 1.0}$	61.5$_{\pm 1.1}$	58.1$_{\pm 1.1}$	57.5$_{\pm 1.9}$	61.1$_{\pm 1.1}$	51.6$_{\pm 0.5}$	53.6$_{\pm 0.9}$	54.1$_{\pm 0.5}$	52.2$_{\pm 0.4}$	52.6$_{\pm 0.9}$	52.6$_{\pm 0.3}$
	SI	52.6$_{\pm 0.0}$	**53.7**$_{\pm 1.3}$	54.5$_{\pm 1.0}$	58.3$_{\pm 2.1}$	58.1$_{\pm 1.1}$	61.6$_{\pm 1.1}$	57.6$_{\pm 0.7}$	57.9$_{\pm 1.7}$	61.6$_{\pm 0.7}$	51.7$_{\pm 0.1}$	51.9$_{\pm 0.7}$	53.6$_{\pm 0.8}$	52.1$_{\pm 0.4}$	52.4$_{\pm 0.8}$	52.7$_{\pm 0.2}$
Rehearsal	AGEM	52.3$_{\pm 0.3}$	52.5$_{\pm 0.1}$	56.1$_{\pm 1.7}$	57.3$_{\pm 1.7}$	57.6$_{\pm 0.9}$	**62.9**$_{\pm 1.5}$	**59.5**$_{\pm 1.8}$	57.3$_{\pm 3.4}$	**63.3**$_{\pm 0.6}$	51.9$_{\pm 0.4}$	53.8$_{\pm 1.5}$	**56.1**$_{\pm 0.9}$	52.2$_{\pm 0.4}$	52.5$_{\pm 0.8}$	52.9$_{\pm 0.4}$
	GEM	**54.8**$_{\pm 1.5}$	50.5$_{\pm 1.0}$	**56.9**$_{\pm 1.3}$	58.2$_{\pm 1.4}$	58.9$_{\pm 1.6}$	61.0$_{\pm 0.8}$	57.9$_{\pm 1.8}$	**58.9**$_{\pm 0.9}$	59.3$_{\pm 1.0}$	**53.0**$_{\pm 0.3}$	54.3$_{\pm 0.6}$	55.4$_{\pm 0.7}$	**54.0**$_{\pm 1.1}$	**55.5**$_{\pm 1.3}$	**58.1**$_{\pm 1.1}$
	Replay	54.4$_{\pm 1.3}$	53.2$_{\pm 1.2}$	55.8$_{\pm 1.9}$	**58.8**$_{\pm 1.4}$	**59.7**$_{\pm 0.4}$	62.1$_{\pm 2.3}$	57.5$_{\pm 1.2}$	56.9$_{\pm 1.1}$	59.9$_{\pm 1.8}$	52.5$_{\pm 0.3}$	53.0$_{\pm 0.6}$	53.8$_{\pm 0.9}$	52.8$_{\pm 1.0}$	52.9$_{\pm 0.6}$	52.7$_{\pm 0.1}$

TABLE III: Final average balanced accuracy for 48hr mortality prediction across Age, Ethnicity, Ward, and Season shift (top), and Hospital shift (bottom). Average performance over 5 runs are presented with bootstrapped 95% confidence intervals. Bold values refer to the best average performance for each model and experiment. For the hospital experiment we report the current performance after training on n hospitals for $n \in \{7, 14, 21, 28, 35\}$. Bracketed numbers refer to the number of different hospitals sequentially trained on thus far.

IV. Results

We present the results of the Domain Incremental experiments in Table III. Results show the final average test balanced accuracy across all tasks for each method. Reported values are means over 5 runs from random initialisation, with bootstrapped 95% confidence intervals. For the Hospital domain shift experiments we present the average performance on all tasks thus-seen as the number of tasks increases (i.e. as the models encounter an increasing number of hospitals).

A. Model Architectures

Models are generally comparable over a small but constant number (40) of training epochs per domain shift, with the exception of Transformers which demonstrated much more volatile performance over repeated runs. Highest training efficiency (measured by number of training epochs required to saturate the current task's loss) was achieved by MLP, followed by LSTM. However a higher training efficiency was correlated with faster and greater forgetting upon introducing new tasks.

B. Continual Learning strategies

Regularization methods showed superior or comparable performance with replay based methods across limited number of domain shifts (Age, Ward, and Ethnicity (broad), Table III top), but decreasing performance as the number of tasks grew large. LwF achieved superior performance on the largest amount of experiments, achieving the lowest degrees of forgetting. For the Hospital domain shift experiments, regularization methods failed to mitigate catastrophic forgetting for n tasks ≥ 5, performing on par with Naive fine tuning (no statistically significant difference in final performances). Such performance is expected of regularization methods on domain incremental problems, having been observed in toy problems generally [38, 17], and in recurrent networks specifically [10]. This is likely due to regularization methods only 'delaying the inevitable' when faced with a large number of tasks, as model parameters are walled off into shrinking locally optimal regions.

Rehearsal methods outperformed all other strategies for a large number of domain shifts. This is consistent with class- and domain-incremental results in other benchmarks [19]. Rehearsal methods all improved with larger storage capacity.

As seen in Table III, regularization methods were generally volatile across a large number of domain shifts, likely corresponding to sets of hospitals more or less similar to the first few encountered. Contrary to this, the rehearsal methods A-GEM and GEM showed relatively stable performance as more hospitals were encountered. This stability in performance over domain shifts demonstrates sustained generalisation as the task population becomes more heterogeneous.

V. Discussion

Our experiments show that simple deep neural networks trained on rich multi-variate sequential data are also prone to catastrophic forgetting in a domain incremental setting.

We observe that regularization methods are prone to more forgetting than rehearsal based methods across a large sequence of tasks, but for few tasks achieve superior or comparable performance to replay based methods (given a fixed small replay buffer).

In the case of patient health records, data may comprise sensitive patient data and hence sharing between institutions or storage over time may require data sharing agreements and ethical approval. This may be prohibitively time-consuming or infeasible, making rehearsal based methods inapplicable. Data-free rehearsal methods such as generative models overcome this issue, but there is a high computational burden to training accurate generative models for such time-series data.

VI. Acknowledgements

Jacob Armstrong is supported by the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1).

References

[1] E Hope Weissler, Tristan Naumann, Tomas Andersson, Rajesh Ranganath, Olivier Elemento, Yuan Luo, Daniel F Freitag, James Benoit, Michael C Hughes, Faisal Khan, et al. The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22(1):1–15, 2021.

[2] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17(1):1–9, 2019.

[3] Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020.

[4] Cecilia S Lee and Aaron Y Lee. Clinical applications of continual learning machine learning. *The Lancet Digital Health*, 2(6):e279–e281, 2020.

[5] Chaitanya Baweja, Ben Glocker, and Konstantinos Kamnitsas. Towards continual learning in medical imaging. *arXiv preprint arXiv:1811.02496*, 2018.

[6] Matthias Delange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Greg Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[7] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[8] Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Anthony Celi, Omar Badawi, and Roger Mark. eICU collaborative research database. *PhysioNet*, 2009. doi:10.13026/C2WM1R. Version 2.0.

[9] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[10] Andrea Cossu, Antonio Carta, Vincenzo Lomonaco, and Davide Bacciu. Continual learning for recurrent neural networks: an empirical evaluation. *arXiv preprint arXiv:2103.07492*, 2021.

[11] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11263, 2019.

[12] Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOPS: continual learning of physiological signals. *CoRR*, abs/2004.09578, 2020.

[13] Nikhil Churamani, Ozgur Kara, and Hatice Gunes. Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition. *arXiv preprint arXiv:2103.08637*, 2021.

[14] Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Alistair Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *medRxiv*, 2021.

[15] Tiago Alves, Alberto Laender, Adriano Veloso, and Nivio Ziviani. Dynamic prediction of icu mortality risk using domain adaptation. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1328–1336. IEEE, 2018.

[16] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

[17] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[18] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113: 54–71, 2019.

[19] David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30:6467–6476, 2017.

[20] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[21] Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard E Turner, and Mohammad Emtiyaz Khan. Continual deep learning by functional regularisation of memorable past. *arXiv preprint arXiv:2004.14070*, 2020.

[22] Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. *arXiv preprint arXiv:1810.11910*, 2018.

[23] Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

[24] Sebastian Farquhar and Yarin Gal. Towards robust evaluations of continual learning. *arXiv preprint arXiv:1805.09733*, 2018.

[25] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *arXiv preprint arXiv:1705.08690*, 2017.

[26] Benjamin Ehret, Christian Henning, Maria R Cervera, Alexander Meulemans, Johannes von Oswald, and Benjamin F Grewe. Continual learning in recurrent neural networks with hypernetworks. *arXiv preprint arXiv:2006.12109*, 2020.

[27] Yong Luo, Liancheng Yin, Wenchao Bai, and Keming Mao. An appraisal of incremental learning methods. *Entropy*, 22(11):1190, 2020.

[28] Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pages 4528–4537. PMLR, 2018.

[29] Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pages 3987–3995. PMLR, 2017.

[30] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[31] Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. *CoRR*, abs/1812.00420, 2018. URL http://arxiv.org/abs/1812.00420.

[32] Vincenzo Lomonaco, Davide Maltoni, and Lorenzo Pellegrini. Rehearsal-free continual learning over small non-iid batches. In *CVPR Workshops*, pages 989–998, 2020.

[33] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European conference on computer vision*, pages 524–540. Springer, 2020.

[34] S. Tang, P. Davarmanesh, Y. Song, D. Koutra, M. Sjoding, and J. Wiens. Mimic-iii and eicu-crd: Feature representation by fiddle preprocessing (version 1.0.0). *Journal of the American Medical Informatics Association*, 27(12): 1921–1934, 10 2020.

[35] Subhrajit Roy, Diana Mincu, Eric Loreaux, Anne Mottram, Ivan Protsyuk, Natalie Harris, Emily Xue, Jessica Schrouff, Hugh Montgomery, Ali Connell, et al. Multitask prediction of organ dysfunction in the icu using sequential sub-network routing. *Journal of the American Medical Informatics Association*, 28(9):1936–1946, 06 2021.

[36] Steven Hicks, Inga Strüke, Vajira Thambawita, Malek Hammou, Pål Halvorsen, Michael Riegler, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *medRxiv*, 2021.

[37] Shachar Kaufman, Saharon Rosset, Claudia Perlich, and Ori Stitelman. Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(4):1–21, 2012.

[38] Xuejun Han and Yuhong Guo. Continual learning with dual regularizations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 619–634. Springer, 2021.