

# Self-Aware SGD: Reliable Incremental Adaptation Framework for Clinical AI Models

Anshul Thakur , Jacob Armstrong, Alexey Youssef, David Eyre , and David A. Clifton

**Abstract**—Healthcare is dynamic as demographics, diseases, and therapeutics constantly evolve. This dynamic nature induces inevitable distribution shifts in populations targeted by clinical AI models, often rendering them ineffective. Incremental learning provides an effective method of adapting deployed clinical models to accommodate these contemporary distribution shifts. However, since incremental learning involves modifying a deployed or in-use model, it can be considered unreliable as any adverse modification due to maliciously compromised or incorrectly labelled data can make the model unsuitable for the targeted application. This paper introduces *self-aware stochastic gradient descent (SGD)*, an incremental deep learning algorithm that utilises a contextual bandit-like sanity check to only allow reliable modifications to a model. The contextual bandit analyses incremental gradient updates to isolate and filter unreliable gradients. This behaviour allows self-aware SGD to balance incremental training and integrity of a deployed model. Experimental evaluations on the Oxford University Hospital datasets highlight that self-aware SGD can provide reliable incremental updates for overcoming distribution shifts in challenging conditions induced by *label noise*.

**Index Terms**—Distribution shifts, incremental learning, medical informatics, COVID-19.

## I. INTRODUCTION

DIGITAL technologies are driving a revolution in healthcare and creating a landscape where data complexity and quantity are growing exponentially. In this landscape, AI is enabling the development of effective *Software as a Medical Device* (SaMD) and digital health applications [1]. Such applications

Manuscript received 25 August 2022; revised 23 November 2022; accepted 8 January 2023. Date of publication 6 February 2023; date of current version 7 March 2023. The work of Anshul Thakur was supported by the EPSRC Healthcare Technologies Challenge under Award EP/N020774/1. The work of Jacob Armstrong was supported by the EPSRC Center for Doctoral Training in Health Data Science under Grant EP/S02428X/1. This work was supported in part by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), and in part by the InnoHK Project at the Hong Kong Centre for Cerebro-cardiovascular Health Engineering (COCHE). (Corresponding author: Anshul Thakur.)

Anshul Thakur, Jacob Armstrong, Alexey Youssef, and David A. Clifton are with the Department of Engineering Science, University of Oxford, OX1 2JD Oxford, U.K. (e-mail: anshul.thakur@eng.ox.ac.uk; jacob.armstrong@hmc.ox.ac.uk; alexey.youssef@eng.ox.ac.uk; david.clifton@eng.ox.ac.uk).

David Eyre is with the Big Data Institute, University of Oxford, OX1 2JD Oxford, U.K. (e-mail: david.eyre@bdi.ox.ac.uk).

This article has supplementary downloadable material available at <https://doi.org/10.1109/JBHI.2023.3237592>, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2023.3237592

take inputs from medical devices, smartphones, wearables, and healthcare records to provide a wide range of clinical functionalities such as patient management and diagnosis. Modern deep learning has been the cornerstone of this AI revolution, expanding into a range of healthcare applications [2], [3] such as disease diagnosis [4], [5], patient monitoring [6], drug discovery [7], and organ transplant allocation [8].

Most of the prevalent deep learning solutions are static in nature and often become outdated if the characteristics of a targeted population change over time [9]. A significant distribution shift in patient variables can render a deployed model practically ineffective. Such shifts are anticipated in the healthcare domain as target populations, underlying disease epidemiology, and treatment protocols often evolve:

- *Shifts in demographics*: Populations evolve naturally over time due to various factors such as immigration [10], changes in fertility/mortality rates [11] and population aging [12].
- *Shifts in epidemiology*: Changes in risk factors and population behaviours can alter disease epidemiology. For example, increased obesity in England during last two decades is expected to affect the prevalence of other chronic diseases [13].
- *Shifts in pathology*: Disease pathophysiology also changes over time, often due to changes in the risk factors of the disease. For example, the major risk factors for cervical cancer are oncogenic variants of Human Papillomavirus (HPV). Although a significant reduction has been witnessed in HPV due to widespread vaccination, other risk factors such as smoking, genetic predisposition, and sexual factors are still prevalent. As such, the nature of future cases is likely to change due to differences in the driving pathophysiology [14], [15].
- *Shifts in disease management protocols and technology*: Protocols, technologies, and tools for disease management evolve over time [16], [17]. These changes in disease management are driven by improved understanding of the disease, early diagnostic advancements, and progress in therapeutics, including personalised therapy.

COVID-19 pandemic presents a clear case study to analyse the impact of the above factors on a population. We witnessed a demographic shift as higher incidence of COVID-19 shifted from older to younger population after the initial pandemic stages [18]. Changes in the management of COVID-19 patients also accompanied this population shift. As our understanding of COVID-19 improved and better preventative (e.g. vaccines)

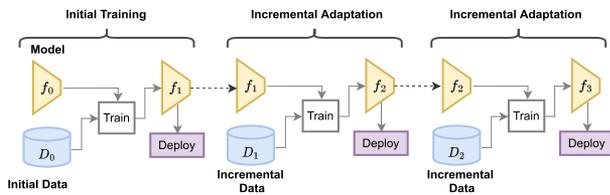


Fig. 1. Illustration of incremental adaptation of the deployed model.

and therapeutic [19] options became available, the protocols were changed to account for these improvements. Another factor altering the nature of COVID-19 is the emergence of new viral variants differing in virulence, pathogenicity and susceptibilities towards vaccines and antivirals [20]. The different outcomes observed in COVID-19 patients over time reflected these trends [21].

As underlying distributions of patient variables shift, models become uncalibrated, and performance degrades over time. A model trained for predicting respiratory deterioration in COVID-19 patients [22], [23] in the early stages of the pandemic may be less effective during the third or fourth wave. Hence, clinical models should be adapted/fine-tuned to stay effective over a prolonged period. Incremental learning [24] provides an effective solution to tackle such distribution shifts, wherein models are iteratively updated to reflect the changes in a target population (Fig. 1). Incrementally updating models over time is a broad topic and has been studied in various contexts over the years, such as *online learning* and *continual learning*. Online learning typically refers to methods for updating models on a potentially infinite incoming stream of training examples. It is generally assumed that all streaming examples belong to the same domain or exhibit no distribution shift [25]. Continual learning also deals with updating models on incoming data over time. However, it iteratively adapts a model to new classes (class-incremental) or domains (domain-incremental) while retaining information about the previously learned classes or domains [25]. Retaining information from previous domains (or previous characteristics of population) is less relevant in the scenario where the target population varies with time, and the updated model is only expected to work on the new population [26]. Hence, this paper mainly deals with such scenarios and considers simple iterative adaptation or fine-tuning of a deployed model to new populations as incremental learning (unless specified otherwise).

Although incremental learning has potential to help in developing dynamic clinical deep learning models, it presents two major challenges to medical practitioners, regulators and AI researchers:

- *Alterations to a deployed model:* Incrementally training a clinical model is precarious as it modifies a deployed or in-use model. An inappropriate modification to critical clinical models, e.g. ICU resource allocation and disease diagnosis models, can hamper performance and may lead to devastating consequences. The primary reason for such performance degradation could be “label noise” that often manifest itself in the incremental data due to labelling errors arising from inaccuracies in diagnosis, coding of diagnoses, and documentation of clinical measurements [27]. Additionally, labels or incremental training data may be

compromised by malicious actors. Since incremental data often arrives in streams or bursts, it is not always feasible to manually check the sanity of labels. If the deployed model is updated with data exhibiting label noise, it may forget the trained task and become unusable.

- *Issues in regulation of iterative AI solutions:* The current regulatory framework of SaMD adopted by healthcare regulators such as the *Medicines & Healthcare products Regulatory Agency* (MHRA) and the *Food and Drug Administration* (FDA) does not enable iterative development approach as it requires products to go through the regulatory certification route once a significant iteration has happened. Recently, FDA launched a new regulatory pathway pilot termed the “PreCert” pathway [28] to regulate the developer of digital health solutions rather than the individual solution itself, facilitating the iterative development. While this is a step towards effective implementation of adaptive digital health solutions, frameworks and tools for quality control over the incremental learning process are still lacking. Quality checks are imperative to ensure that newly incorporated information and updates lead to non-inferior performance of deployed models.

This paper aims to provide a framework for reliable incremental adaptation of clinical deep learning models that addresses the above mentioned challenges to a great extent. To this aim, we introduce *self-aware stochastic gradient descent* (SGD) that acts as a wrapper over standard stochastic gradient descent (SGD) and imposes a sanity check over gradients for reliable incremental learning. In the presence of label noise in incremental data, the proposed algorithm effectively balances the adaptation of a deployed clinical model and maintaining its integrity. *Self-aware SGD* exploits a deep neural network-based contextual bandit that analyses the magnitude and direction of a gradient update to predict its impact on the performance of the deployed model. A gradient update that results in performance deterioration is deemed harmful and filtered out. As a result, *self-aware SGD* assures that a model is updated reliably and robustly.

Apart from standard incremental adaptation, *Self-aware SGD* is also compatible with state-of-the-art (replay-based) domain-incremental learning methods. By preserving the integrity of a deployed model during incremental learning, this paper exhibits a potential mechanism to create new iterative AI solutions for healthcare that asserts quality control over themselves. Tools like the proposed algorithm can potentially be a key component of how regulators control and exert quality assurance over rolled out incremental digital health solutions.

The major contributions of this paper are listed below:

- This paper proposes self-aware SGD, a reliable incremental learning algorithm, that can provide effective incremental adaptation of deployed deep learning models under challenging label noise conditions without compromising their integrity.
- The proposed algorithm provides a blueprint for the possible future studies dealing with clinical incremental learning applications. Unlike existing studies, this paper has highlighted the requirement of reliability and self-quality control in the clinical incremental learning solutions.

- This paper provides the evidence in favor of dynamic clinical models by utilising the Oxford University Hospitals (OUH) data, collected between 2016 and 2021, to show the inappropriateness of static clinical models in handling the evolving target populations.

The rest of this paper is organised as: Section II provides a background of the existing incremental learning frameworks within the context of healthcare informatics. Section III describes the proposed self-aware SGD. Section IV and V describe the experimental setup and analysis of the results, respectively. Finally, Section VI concludes this paper.

## II. BACKGROUND

### A. Incremental Learning

Incremental learning has seen limited exploration in healthcare applications. In [26], the authors proposed an adaptive risk prediction system that detects distribution shifts and adapts the models to cater to the changing population. Guo et al. [29] and Alves et al. [30] also tackled the problem of temporal shifts in the target population. However, instead of incremental learning, these studies considered domain adaptation for handling distribution shifts. In [31], the authors benchmarked state-of-the-art domain-incremental learning algorithms for longitudinal electronic health records or multivariate clinical time-series data and found that replay-based domain-incremental algorithms outperform the other counterparts. On the similar lines, Kiyasseh et al. [32] had explored replay-based domain-incremental learning on physiological signals earlier. Domain-incremental learning has also been used in some studies to overcome distribution shifts in X-Ray images that arise either due to differences in sensors or in the underlying patient population [33], [34].

### B. Robust Deep Learning Training Under Label Noise

Training deep learning models using data with label noise is a well-studied problem in deep learning literature. The efficient noise-robust mechanisms can isolate training signals from overwhelming noise and provide effective training. In [35], the authors proposed a meta-learning framework where a simple one layer MLP is used to learn the importance of each sample in a training dataset. In theory, the importance of wrongly labelled examples will be lower, and they will have no impact on the overall training. Xu et al. [36] proposed an information theoretic loss function that is robust any pattern of label noise. In [37], the authors highlighted *early learning regularisation* to propose a label-noise robust training mechanism. In the early learning phase, the model mostly learns from correctly labelled examples. During the later phase, the model starts learning from noisy labels resulting in feature interference. Building on this observation, the authors proposed a regularisation mechanism that tries to force the model outputs to be consistent with the model predictions obtained in the early training phase.

All the existing noise robust training mechanisms only deal with training models from scratch. These methods do not impose any integrity constraints on a deployed model and may degrade performance after incremental training (see Section V).

### C. Comparison With Self-Aware SGD

Most of the existing incremental studies do not consider the reliability and regulatory concerns posed by modifications to a deployed clinical model during incremental adaptation. Unlike the existing methods, the proposed self-aware SGD is specially designed to strike an effective balance between maintaining the performance of the deployed model and adaptation to the evolving population. To the best of our knowledge, this paper is the first attempt in performing incremental learning under label noise.

## III. PROPOSED SELF-AWARE SGD

This section elaborates the proposed self-aware SGD based incremental learning framework. Here, we first describe the problem statement. Then, we describe the concept of gradient consistency that forms the basic building block of the proposed self-aware SGD. Finally, self-aware SGD is presented.

### A. Problem Statement

A deployed or in-use clinical deep learning model has to be updated using batches of incremental data that are arriving in bursts. The aim is to come up with an incremental learning algorithm that makes sure that the application of incremental updates does not result in any catastrophic drop in performance of the deployed model. This desired algorithm must work under the assumption that some gradient updates are going to be harmful (arising from batches with drastic label noise) for the deployed model while others will help model in adapting itself to the evolving population. It should filter out the harmful updates while applying the required incremental updates to adapt the model.

### B. Gradient Consistency

A deployed model has already been trained for the targeted task. It has learned the features or semantics associated with each class. When this model is incrementally updated using a batch containing incorrectly labelled data, we are forcing the model to switch this semantic-class relationship. For example, we are dealing with a model deployed for respiratory deterioration prediction and the lower SPO<sub>2</sub> is associated with respiratory deterioration. However, the gradient update computed from a wrongly labelled batch may force the model to associate lower SPO<sub>2</sub> with no respiratory deterioration event. This implies that the incremental gradient update is not *consistent* with the historical gradient updates that have been used to train the deployed model.

All historical gradient updates (that resulted in current state of the deployed model) can be summarised as change in parameters from their initial random state to the current state. Suppose  $\theta_0$  and  $\theta_1$  represent the random and current state of model parameters. Then, the historical gradient update can be defined as:

$$\mathbf{g}_h = -(\theta_1 - \theta_0) = \theta_0 - \theta_1 \quad (1)$$

**Algorithm 1:** Process of Training a Contextual Bandit Model  $B_\phi()$  in *Self-Aware SGD*.

---

```

1: INPUTS  $\leftarrow f_\theta$ : model,  $\mathbf{g}_h$ : historical grad,  $B_\phi$ : bandit,
    $\mathcal{D}_{inc}, \mathcal{D}_{val}$ : Incremental & Validation Data,  $\tau$ : Threshold
2: OUTPUT  $\leftarrow \phi$ , trained bandit parameters
3:  $\mathcal{P} = [], \mathcal{R} = []$ 
4:  $R_{prev} = \text{SCORE}(\mathcal{D}_{val}, f_\theta()) \triangleright$  Compute AUROC score
5:  $\mathcal{B} \leftarrow \text{SAMPLE-BATCHES}(\mathcal{D}_{inc}) \triangleright |\mathcal{B}|$  batches
// Obtaining rewards and features for training  $B_\phi()$ 
6: for all  $(\mathbf{b}, \mathbf{l}) \in \mathcal{B}$  do
7:    $L = \mathcal{L}_{CE}(f_\theta(\mathbf{b}), \mathbf{l}) \triangleright$  Cross-entropy loss
8:    $\mathbf{G}_\theta = \nabla_\theta L \triangleright$  Gradient
9:    $N = \|\mathbf{G}_\theta\|_2 \triangleright$  Compute norm
10:   $C = \text{COSINE-SIMILARITY}(\mathbf{g}_h, \mathbf{G}_\theta)$ 
11:   $\theta' = \theta - \eta \mathbf{G}_\theta \triangleright$  Gradient update with learning rate  $\eta$ 
12:   $R_{cur} = \text{SCORE}(\mathcal{D}_{val}, f_{\theta'}())$ 
13:   $r = R_{cur} - R_{prev}$ 
14:  if  $r > \tau$  then
15:     $\mathcal{R}.\text{APPEND}(1)$ 
16:     $\mathcal{P}.\text{APPEND}([N, C])$ 
17:  else
18:     $\mathcal{R}.\text{APPEND}(0)$ 
19:     $\mathcal{P}.\text{APPEND}([N, C])$ 
// Training  $B_\phi()$ 
20:  $\mathcal{C} \leftarrow \text{SAMPLE-BATCHES}(\mathcal{P}, \mathcal{R}) \triangleright |\mathcal{C}|$  batches
21: for all  $(\mathbf{c}, \mathbf{l}) \in \mathcal{C}$  do
22:    $L = \mathcal{L}_{MAE}(B_\phi(\mathbf{c}), \mathbf{l}) \triangleright$  MAE loss
23:    $\mathbf{G}_\phi = \nabla_\phi L \triangleright$  Gradient
24:    $\phi' = \phi - \alpha \nabla_\phi L$ 
25:    $\phi = \phi' \triangleright$  Update  $\phi$ 
26: Return  $\phi$  or  $B_\phi()$ 

```

---

At the current state of model parameters ( $\theta_1$ ), an incremental gradient update,  $\nabla_{\theta_1}$ , must be consistent with  $\mathbf{g}_h$  to preserve the integrity of the model. This consistency can be defined as:

- *Gradient norm*: The norm of a consistent gradient update must be lesser than an inconsistent one. The lower norm implies that the magnitude of gradient is lower, and it won't cause any significant changes to the current model state [38].
- *Cosine similarity*: The higher cosine similarity between  $\mathbf{g}_h$  and  $\nabla_{\theta_1}$  implies that  $\nabla_{\theta_1}$  is going to update the parameters in a direction similar to historical parameter updating. Hence,  $\nabla_{\theta_1}$  can be considered as consistent. The opposite is true if cosine similarity is lower. Since gradients are tensors (not vectors), we vectorise  $\nabla_{\theta_1}$  and  $\mathbf{g}_h$  to compute cosine similarity:

$$C_{\mathbf{g}_h, \nabla_{\theta_1}} = \frac{\nabla_{\theta_1} \cdot \mathbf{g}_h}{\|\nabla_{\theta_1}\| \times \|\mathbf{g}_h\|}. \quad (2)$$

To visualise gradient consistency in action, we computed gradients using batches with correctly and incorrectly labelled examples for a model trained/deployed for respiratory deterioration prediction (see Section IV). We corrupted batches with four different label noise probabilities, i.e. we flipped the label of each

**Algorithm 2:** Reliable Incremental Adaptation of Deployed Model Using *Self-Aware SGD*.

---

```

1: INPUTS  $\leftarrow f_\theta$ : model,  $\mathbf{g}_h$ : historical grad &  $B_\phi$ : bandit
2: OUTPUT  $\leftarrow \theta$ , trained model parameters
3: while  $\mathcal{D}_{inc}$  arrives do  $\triangleright$  Incremental data
4:    $B_\phi \leftarrow \text{TRAIN-BANDIT}(B_\phi, f_\theta, \mathcal{D}_{inc}) \triangleright$  Algorithm 1
5:    $\theta_0 = \theta \triangleright$  Copy of input model parameters
6:    $\mathcal{B} \leftarrow \text{SAMPLE-BATCHES}(\mathcal{D}_{inc}) \triangleright |\mathcal{B}|$  batches
7:   for all  $(\mathbf{b}, \mathbf{l}) \in \mathcal{B}$  do
8:      $L = \mathcal{L}_{CE}(f_\theta(\mathbf{b}), \mathbf{l}) \triangleright$  Cross-entropy loss
9:      $\mathbf{G}_\theta = \nabla_\theta L \triangleright$  Gradient
10:     $N = \|\mathbf{G}_\theta\|_2 \triangleright$  Gradient norm
11:     $C = \text{COSINE-SIMILARITY}(\mathbf{g}_h, \mathbf{G}_\theta)$ 
12:     $P = B_\phi([N, C]) \triangleright$  Bandit Prediction
13:    if  $P < 0.5$  then
14:      skip remaining steps
15:    else  $\theta' = \theta - \alpha \nabla_\theta L$ 
16:       $\mathbf{g}_h = \theta_0 - \theta' \triangleright$  Update historic gradient
17:       $\theta = \theta' \triangleright$  Update  $\theta$ 
18: Return  $\theta$  or  $f_\theta()$ 

```

---

example (in each batch) with a uniform random probability of 0.2, 0.4, 0.6 and 0.8. Fig. 2(a) and (b) depicts the difference between gradient norm and cosine similarity of gradients computed from corrupted and ‘‘pure’’ batches. The analysis of this figure highlights a clear difference between correctly labelled batches and batches with label noise. Hence, both these properties can be used to create an automatic sanity check to identify and reject the inconsistent gradients.

### C. Self-Aware SGD

Self-aware SGD can be seen as combination of two stages: training the bandit model and performing reliable incremental learning using the trained bandit model. Following the nomenclature from the contextual bandit literature, these stages can be seen as analogous to *exploration* and *exploitation*. In the exploration phase, an agent learns to associate the actions and contexts with rewards. Hence, it can be considered as a training phase. On the other hand, an agent exploits the accumulated knowledge to take actions and obtain rewards in the exploitation phase. The details of these two stages with respect to incremental learning are discussed below:

- *Training bandit model*: The bandit model is trained or updated before each bout of incremental learning. As incremental data arrives in bursts of a few batches at a time, only these batches are analysed to update or train the bandit model. Since there are no labels (or rewards) to train the bandit model, the first step is to analyse gradient updates and their impact on the performance of the initial or deployed model over validation examples. The gradient updates are computed for each incremental batch, and each gradient update is processed to obtain its norm and cosine similarity with historical model updates. These gradient norms and cosine similarities are accumulated and are used as input features (context) to train the bandit model.

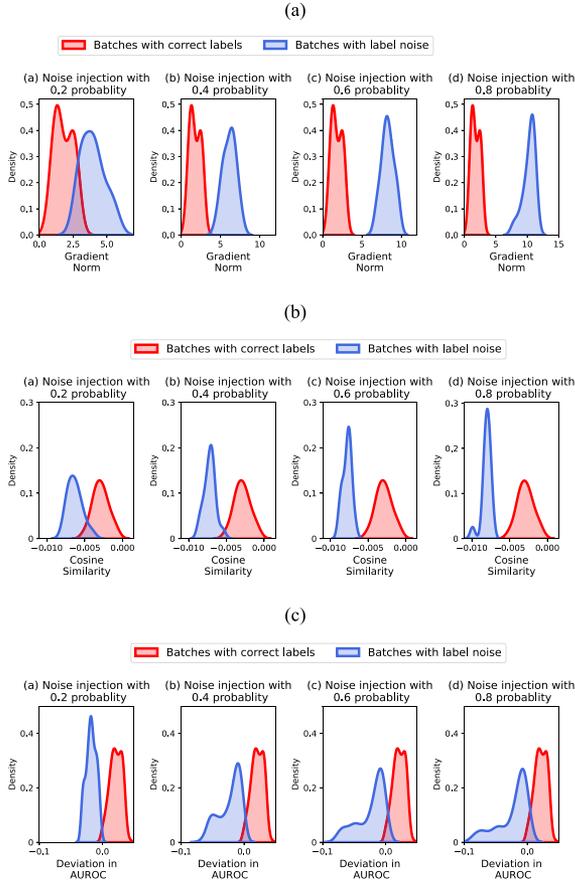


Fig. 2. Difference between distribution of (a) norm and (b) cosine similarity of gradients computed from batches with correct and incorrect labels. (c) Illustration of differences in distribution of *deviation in AUROC* induced by incremental adaptation with correctly and incorrectly labelled batches. Kernel density estimation (with Gaussian kernel) is used to estimate these distributions.

The labels or rewards are generated for each batch or its corresponding gradient without affecting the initial model. For each gradient update, a copy of the initial model is created and the gradient update is applied to this copy. Then, we compute performance metrics such as the area under the ROC curve (AUROC) on the validation examples using the “updated copy” model. Based on the performance of the initial and updated model, we define reward  $R$  as:

$$R = \begin{cases} 1, & \text{if } (r_\theta - r_{\theta'}) > \tau \\ 0, & \text{if } (r_\theta - r_{\theta'}) < -\tau. \end{cases} \quad (3)$$

Here  $r_\theta$  and  $r_{\theta'}$  represent the AUROC obtained by the deployed/initial model and incrementally updated copy of the initial model, respectively.  $\tau$  is a user-defined parameter used to reject the confusing gradients from this training process. In terms of distribution, these confusing gradients belong to the region overlapped by both correctly and incorrectly labelled batches (see Fig. 2(c)). After computing gradient properties and their corresponding rewards or labels, the bandit model is trained to predict whether an input gradient is consistent or inconsistent with the initial

model. An update is regarded as inconsistent if it results in performance deterioration and vice-versa.

- *Reliable incremental adaptation using trained bandit model*: Once the bandit model has been trained, it can be deployed as a wrapper over the standard gradient descent (or any of its variants). During incremental training, the gradient is computed for a batch as in any standard deep learning framework. Then, this gradient is processed to compute its norm and cosine similarity with respect to historical gradient updates (as discussed earlier). The contextual bandit model intakes these gradient properties and predicts the gradient consistency. The gradient update is only applied if it is deemed consistent, and hence, the integrity of the deployed is preserved.

*Implementation details*: Suppose  $f_\theta()$  be an initial/deployed model, with  $\theta$  defining its parameters or weights. Similarly, let  $B_\phi()$  be a DNN acting as an agent or a bandit model. First,  $B_\phi()$  is trained using the set of labelled incremental batches  $\mathcal{X}_{inc}$ . Then, *self-aware SGD* utilises the trained bandit model,  $B_\phi()$ , to identify the consistent gradient updates for adapting the initial model  $f_\theta()$ . Algorithm 1 documents the process of training the bandit model and Algorithm 2 illustrates the process of deploying this bandit model for incremental learning.

Since the proposed algorithm is aware of the historical gradient direction and determines the nature of the gradient updates without any external stimulus, it has been referred to as *self-aware SGD* in this paper. This algorithm works on the assumption that incremental data has a mix of corrupted and pure batches. In real life, we may encounter scenarios where all incremental batches are pure (best-case scenario), or all are corrupted (worst-case scenario). We can borrow a handful of batches from validation data to handle such cases. Since batches from validation data are pure, we can add artificial label noise to create their corrupted versions. These noisy and original validation examples can be added to the incremental data for the training bandit model.

#### D. Extending Self-Aware SGD for Domain-Incremental Learning

In domain-incremental learning, a model must retain information about the previous domains (or populations) while adapting to new domains. *Self-aware SGD* is consistent with replay-based domain-incremental methods (which are considered state-of-the-art). In replay-based methods, a subset of previous domain examples is mixed with new domain data during incremental training. This subset of previous domain examples *replays* the information in the corresponding domain to the model and avoids catastrophic forgetting. Since *self-aware SGD* identifies and isolates the inconsistent gradients, the corresponding training batches can be removed from contention of being utilised in any replay-based method. Hence, the process of selecting examples for replays is limited to the examples considered consistent by self-aware SGD. Since all replay-based methods differ only in storing and retrieving the previous domain examples, any replay-based method can be augmented with self-aware SGD effectively.

TABLE I

NUMBER OF PATIENTS AND EXAMPLES IN EACH SUB-DATASET OBTAINED FROM OXFORD UNIVERSITY HOSPITALS (OUH) DATA

Dataset	# Patients	# Examples	# Respiratory Deterioration Events
2016	54,812	1,439,950	26,827 (1.86 %)
2017	64,103	1,723,223	25,634 (1.49 %)
2018	64,693	1,696,403	23,194 (1.37 %)
2019	70,919	1,988,742	23,203 (1.17 %)
COVID-1	485	17,756	141 (0.79 %)
COVID-2	1,730	59,956	460 (0.77 %)

#### IV. EXPERIMENTS

This section describes the dataset and experiments designed to evaluate the performance of *self-aware SGD* and *self-aware SGD with replays*.<sup>1</sup>

##### A. Dataset Used

Patient records from the *Infections in Oxfordshire Research Database (IORD)* are used for evaluating the proposed framework.<sup>2</sup> This data is collected from patients admitted to Oxford University hospitals (OUH) between January 2016 and June 2021. Patients admitted between January 2016 and December 2019 exhibited various underlying conditions such as pneumonia, heart failure, and asthma. In contrast, the data between March 2020 and June 2021 is only collected from patients with PCR confirmed COVID-19. To simulate an incremental learning setup, we temporally divide the data into six subsets: *2016 dataset*, *2017 dataset*, *2018 dataset*, *2019 dataset*, first COVID-19 dataset (March 2020 to July 2020), and second COVID-19 dataset (August 2020 to June 2021) dataset. The first COVID-19 dataset (*COVID-1*) corresponds to first COVID-19 wave, whereas the second COVID-19 dataset (*COVID-2*) corresponds to second and third waves.

Patient features are sampled at irregular time intervals reflecting ad hoc clinical measurements taken by hospital staff. Each sample is characterised by a 77-dimensional feature vector and a binary label (retrospectively generated) signifying *respiratory deterioration* within the next 24 hours [23]. Features include demographic characteristics, vital sign measurements, laboratory test results, and inspired oxygen concentration ( $\text{FiO}_2$ ). The detailed information regarding these features and data pre-processing can be found in [23] and in the supplementary document.

Table I documents the number of patients, number of samples, and percentage of samples exhibiting respiratory deterioration in each sub-dataset after pre-processing.

<sup>1</sup>More experiments dealing with EHR time-series data are presented in the supplementary document.

<sup>2</sup>Dataset is not publicly available as it contains personal/sensitive patient information. However, it can be obtained from the Infections in Oxfordshire Research Database (IORD), subject to an application and research proposal meeting on the ethical and governance requirements of the database.

##### B. Designed Experiments

We train a deep neural network for the task of respiratory deterioration prediction in an incremental learning setup depicted in Fig. 3. Following experiments were designed to evaluate the performance of self-aware SGD using this setup:

- *Self-aware SGD vs. standard SGD for incremental adaptation*: The performance of self-aware SGD is compared against the standard SGD (*normal training*) in presence of the label noise for incremental adaptation or fine-tuning. The noisy conditions are simulated by randomly flipping labels of the incremental training batches. A training batch is selected with probability  $b$ , and the label of each example in a selected batch is flipped with probability  $p$ . We refer to  $b$  and  $p$  as *batch* and *example* probabilities, respectively. We used batch probabilities of 0.5 and 0.75, and a fixed example probability of 0.8 i.e. we randomly corrupt labels of approximately 50% and 75% of available batches. In each corrupted batch, labels of approximately 80% of randomly chosen examples are flipped.
- *Comparison against existing noise robust algorithms*: In this experiment, the performance of self-aware SGD is compared against the well-known noise-tolerant DNN training frameworks. These methods include determinant-based mutual information (DMI) loss function [36], early learning regularisation (ELR) [37] and meta-weight net [35] (discussed in Section II).
- *Self-aware SGD with replays for domain-incremental learning*: The ability of *self-aware SGD with replays* (Section III) for domain-incremental learning is evaluated. For this experiment, we augment *self-aware SGD* with *greedy sampler and dumb learner* (GDumb) [39], a simple replay-based domain-incremental learning method. GDumb selects a fixed number of most recently encountered examples from each domain or population and stores them in a fixed-size memory buffer that is equally divided among all populations. In this setup, *self-aware SGD* only stores the most recently encountered training batches in memory buffer of 50,000 examples if their gradients are deemed consistent. During incremental adaptation, *self-aware SGD* queries the memory buffer to obtain the examples corresponding to previous populations. These examples are mixed with new incremental data for reliable model adaptation.

We incrementally trained the model using *self-aware SGD with replays* in a label noise setup where the batch probability of 0.5 is used for label corruption.

##### C. Models and Parameter Setting

A fully-connected neural network or DNN is trained for predicting respiratory deterioration events. The model consists of three fully-connected or dense layers having 308 ( $77 \times 4$ ), 231 ( $77 \times 3$ ) and 1 hidden units. The first two dense layers are followed by the rectified linear activation function, and the last layer is followed sigmoid activation function. A dropout of 0.25 is used between dense layers to regularise the model.

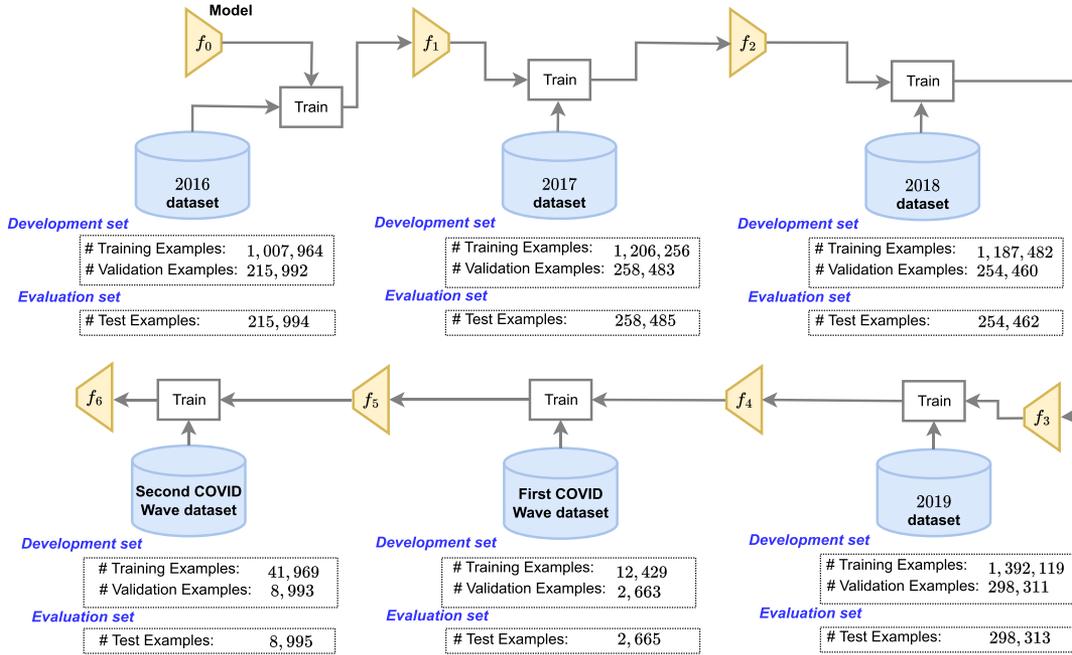


Fig. 3. Incremental learning framework used in the experimental evaluation of self-aware SGD. This illustration exhibits the order of datasets used for incremental model adaptation/fine-tuning for the task of *respiratory deterioration prediction*. The numbers of training, validation, and test examples in each dataset are also depicted (70%, 15%, and 15% of the total examples respectively). Patients sampled for training, testing, and validation are different.

Binary cross-entropy is used as a loss function, and we use standard SGD with 0.9 momentum and 0.0001 learning rate as an optimiser to train the initial model on *2016 dataset* (see Fig. 3). A batch size of 512 examples is used across all experiments. Also, we train each model (initial or incremental) for 100 epochs and use early stopping to store only the best performing version of the model. During incremental adaptation using self-aware SGD, we used also use SGD optimiser with 0.9 momentum and 0.0001 learning rate.

We use a three-layered DNN consisting of layers with 8 nodes, 4 nodes and 1 node as a contextual bandit model. The first two nodes are followed by rectified linear activation whereas the last node is followed by sigmoid activation function. The loss function used to train this model is mean average error (MAE) [40], and Adam with a learning rate of 0.0001 is used as optimiser. We used a batch size of 128 across to train bandit model across all experiments.

For comparative analysis, we train DNN with determinant-based mutual information (DMI) loss function [36] using SGD optimiser with 0.0001 learning rate and 0.9 momentum. In ELR [37], we used a regularisation coefficient of 7 ( $r = 7$ ) and weight factor of 0.7 ( $t = 0.7$ ) for generating pseudo-labels. SGD with a learning rate of 0.001 and momentum of 0.9 was used as an optimiser. For meta-weight net, we have used SGD with a 0.001 learning rate and 0.9 momentum as an optimiser to train the respiratory deterioration prediction model. To train the weighing MLP, SGD with 0.001 and zero momentum was used. All these parameters are chosen using hyperparameter tuning on validation examples.

## V. RESULTS AND DISCUSSION

### A. Self-Aware SGD vs. Standard SGD for Incremental Learning

Fig. 4 depicts the performance of *self-aware SGD* and standard incremental training (*normal training*) within the incremental framework illustrated in Fig. 3. The analysis of Fig. 4 highlights the following:

- In standard incremental training (*normal training*), the presence of label noise in the incremental data results in a relative drop of  $24.4(\pm 1.45)\%$  at  $b = 0.5$  and  $44.4(\pm 2.1)\%$  at  $b = 0.75$  in the performance of the deployed model on *2016 dataset* (Fig. 4(a)). This drop upholds our claim of the requirement of reliability in clinical incremental learning.
- The performance of *self-aware SGD* is near-optimum across all datasets, and no performance drop is observed in any of the noisy conditions unlike *normal training*. This shows that *self-aware SGD* can effectively identify and filter out “harmful” gradient updates during incremental training, preserving the integrity of the deployed model.
- The analysis of Fig. 4(b) and (c) shows that the model trained on *2016 dataset* (initial model) exhibits excellent performance on the *2017* and *2018* datasets. Incremental learning using *self-aware SGD* (or standard training in no label-noise setup) with *2017* and *2018* data does not result in any noticeable improvement. This shows that the nature of *2016*, *2017* and *2018* datasets is similar,

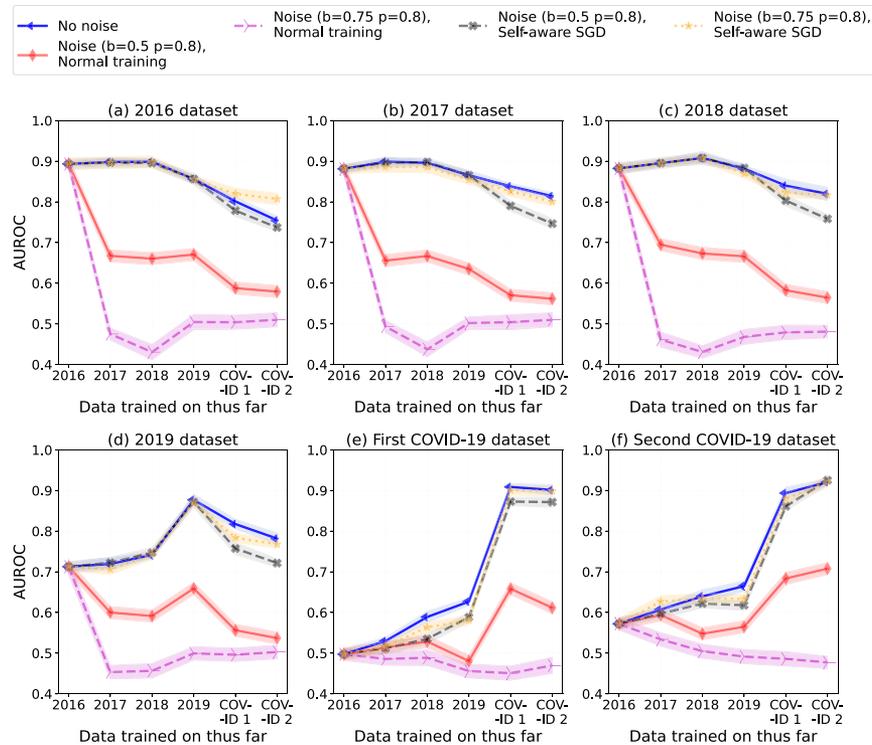


Fig. 4. Performance comparison between the normal DNN training and *self-aware* SGD under simulated label noise conditions on the test examples from (a) 2016, (b) 2017, (c) 2018, (d) 2019, (e) first, and (f) second COVID-19 datasets. The baseline is normal DNN training under no label noise. Each x-axis point illustrates the datasets that have been used to update the model till that point. Note we display performance on previous years' test sets after incremental training on subsequent years to evaluate the degree of catastrophic forgetting.

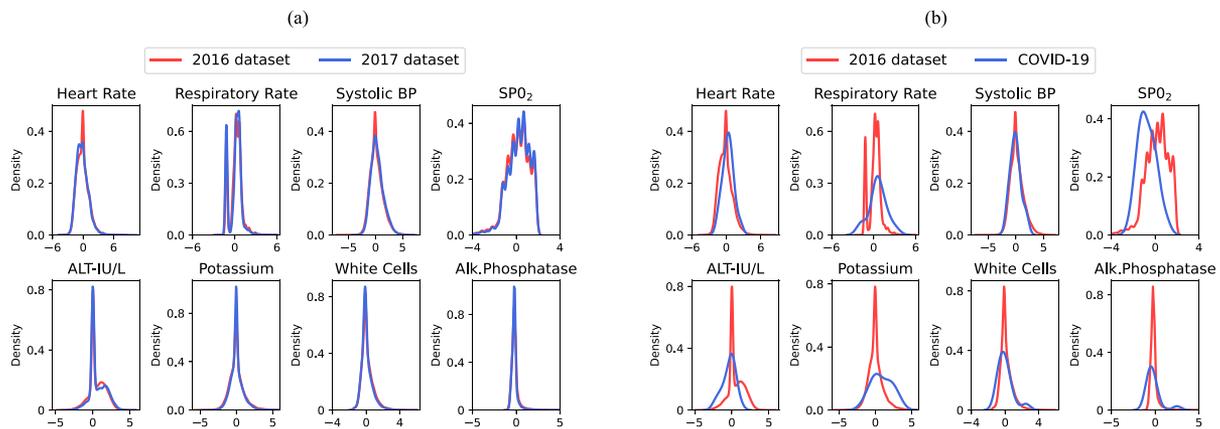


Fig. 5. An illustration of distribution shifts in 8 randomly selected features of (a) 2016 and 2017 datasets, and (b) 2016 and first COVID-19 wave datasets. Kernel density estimation (KDE) with Gaussian kernel is used for estimating these feature distributions.

and there is little to no distribution shift between these datasets. This is illustrated in Fig. 5(a) showing the empirical distributions of 8 different features from the 2016 and 2017 training datasets. However, the model trained on 2016 or incrementally updated using 2017 and 2018 datasets demonstrates high but sub-optimal performance on 2019 test examples (Fig. 4(d)). Incrementally updating using 2019 training examples results in a significant performance improvement. This highlights that the features of 2019 and previous datasets exhibit a minute distribution

shift such that the previous model is still effective but not optimum.

- A more significant change is observed between COVID-19 and the other datasets. The initial model or incrementally adapted model using 2017, 2018 and 2019 datasets is practically unusable on both COVID-19 datasets (Fig. 4(e) and (f)). However, after incremental adaptation using COVID-19 data, the performance on the COVID-19 datasets is significantly improved. This signifies a distribution shift between historical and COVID-19 data (see Fig. 5(b)).

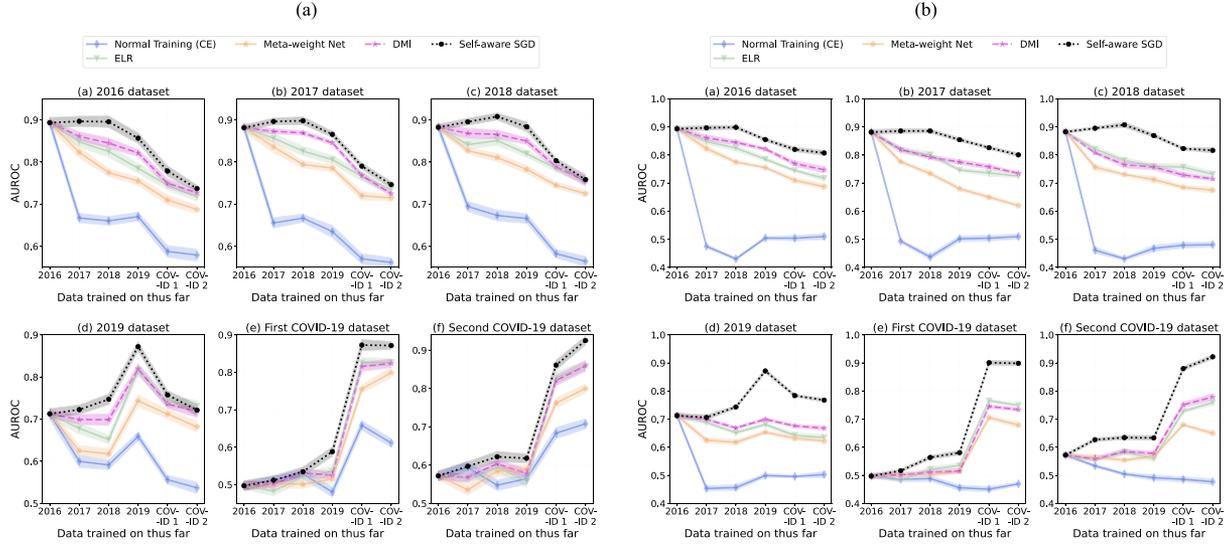


Fig. 6. Performance comparison between *self-aware SGD* and the comparative methods under simulated label noise defined by batch probabilities of (a) 0.5 and (b) 0.75. The comparative methods include DMI loss function [36], early learning regularisation (ELR) [37] and Meta-weight net [35].

- The improvement in performance after incremental adaptation using *self-aware SGD* (Fig. 4(d), (e) and (f)) under noisy label conditions highlights that it not only preserves the integrity of the model but also allows an effective incremental adaptation to overcome the distribution shifts induced by evolution of the underlying population.
- The paired t-tests are performed to analyse the statistical significance of the performance improvement in *self-aware SGD* over *normal training* across all experiments under label noise. The null hypothesis is that there is no statistical difference in scores, and we reject null hypothesis if  $p < 0.005$ . Apart from COVID datasets, the performance improvement achieved by the proposed method is statistically significant. In case of COVID datasets, the significant distribution shifts have rendered the models trained using 2016 to 2018 datasets ineffective (even in no noise scenarios). However, after adapting models to COVID datasets, we again witness a significant improvement in the performance of *self-aware SGD*.

## B. Comparison Against Existing Label-Noise Robust Deep Learning Methods

We compare the performance of *self-aware SGD* with existing label-noise robust deep learning methods under the simulated label-noise configurations. Fig. 6(a) and (b) depict the results of this comparison at batch probabilities of 0.5 and 0.75, respectively. The analysis of these figures highlight that *self-aware SGD* significantly outperforms existing methods in most experimental setups. The paired t-tests between scores obtained by *self-aware SGD* and the comparative methods at batch probability of 0.75 highlight that the performance improvement by *self-aware SGD* is statistically significant in approximately all scenarios ( $p < 0.005$ ). At a batch probability of 0.5, the performance of *self-aware SGD* and DMI loss is statistically comparable across many experimental setups ( $p > 0.005$ ). In spite of that,

*self-aware SGD* exhibits noticeable improvement over the other comparative methods.

Although the existing methods exhibit better performance than *normal training*, they fail to preserve the integrity of deployed models and result in performance deterioration. This behaviour is expected as none of these methods are designed for incremental learning. Although these methods can isolate the relevant training signals from the noisy labels with some success, they impose no constraint on the preserving historical performance of the deployed models.

## C. Catastrophic Forgetting During Incremental Learning

On analysing the performance of *normal training* under no label noise in Fig. 4(a), it is clear that catastrophic forgetting is not observed on incremental adaptation of model using 2017 and 2018 data. However, a relative drop of 9.88% and 15% is observed after incremental training on *COVID-1* and *COVID-2* datasets. This drop is due to the distribution shift between earlier datasets (2016–2018) and COVID-19 datasets. Incremental training on new datasets interferes with previously learned features and results in this performance drop. *Self-aware SGD* also exhibits catastrophic forgetting. However, it undergoes lesser adaptation and exhibits less catastrophic forgetting than standard training as most of the noisy incremental data is rejected. Similar catastrophic behaviour is observed in Fig. 4(b), (c) and (d) after adaptation with the COVID-19 datasets.

## D. Self-Aware SGD With Replays for Domain Incremental Learning

The results of this experiment are depicted in Fig. 7. The analysis of this figure highlights that *self-aware SGD with replays* avoids the catastrophic forgetting witnessed in *self-aware SGD*. Incremental training with COVID-19 datasets does not cause any significant drop in performance over 2016 to 2019 datasets

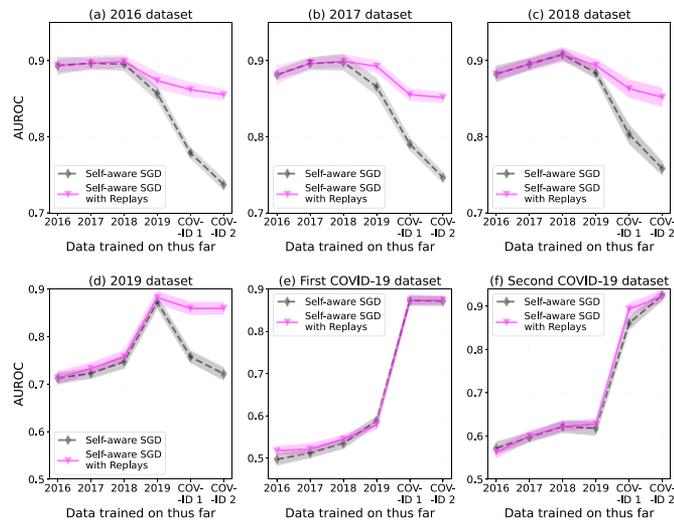


Fig. 7. Performance comparison between *self-aware SGD* and its variant with replays to avoid catastrophic forgetting.

(Fig. 7(a)–(d)). Hence, *self-aware SGD with replays* can overcome both label noise and catastrophic forgetting challenges.

## VI. CONCLUSION

This paper highlighted that clinical models must evolve with time to match the dynamic nature of diseases and target populations. A traditional static model cannot cope with the distribution shifts seen in the underlying clinical variables as the nature of population health and diseases change, and hence become ineffective over time. This paper further argued that incremental learning may help in coming up with dynamic clinical AI solutions to tackle evolving populations while bring the reliability and regulatory concerns in incremental learning to the foreground. To alleviate these concerns, this paper conceptualised an incremental learning framework, i.e. *self-aware SGD*, with a sanity check over gradient updates to allow only reliable changes to a deployed model. The experimental results highlight that these sanity checks can indeed allow an incremental learning framework to preserve the integrity of deployed models even in the presence of extreme label noise. Hence, *self-aware SGD* and similar future algorithms can pave the way for exploiting incremental learning to develop reliable SaMD solutions.

In comparison to the standard SGD, *self-aware SGD* is computationally expensive as it requires a trained bandit model. However, this computational overhead empowers the proposed framework to tackle the challenging label-noise conditions. Apart from that, a major limitation of the current version of *self-aware SGD* is that it has only targeted the prediction tasks (binary classification problems). Although the prediction tasks form a bulk of automated decision support systems, it would be beneficial to extend the proposed framework to more generic and challenging settings such as multi-class classifications and temporal segmentation. Future work will deal with extending this work to these use-cases using advanced reinforcement learning algorithms.

## ACKNOWLEDGMENT

DE is a Big Data Institute Robertson Fellow. DAC is an Investigator in the Oxford Suzhou Centre for Advanced Research (OSCAR), Suzhou, China, and in the Pandemic Sciences Institute, University of Oxford, Oxford, U.K. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, the Department of Health, InnoHK – ITC, or the University of Oxford.

## REFERENCES

- [1] B. Allen, “The role of the FDA in ensuring the safety and efficacy of artificial intelligence software and devices,” *J. Amer. College Radiol.*, vol. 16, no. 2, pp. 208–210, 2019.
- [2] D. Ravi et al., “Deep learning for health informatics,” *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.
- [3] A. Thakur, P. Sharma, and D. A. Clifton, “Dynamic neural graphs based federated reptile for semi-supervised multi-tasking in healthcare applications,” *IEEE J. Biomed. Health Informat.*, vol. 26, no. 4, pp. 1761–1772, Apr. 2021.
- [4] J. Diaz-Escobar et al., “Deep-learning based detection of covid-19 using lung ultrasound imagery,” *PLoS One*, vol. 16, no. 8, 2021, Art. no. e0255886.
- [5] L. Jin et al., “Deep learning extended depth-of-field microscope for fast and slide-free histology,” *Proc. Nat. Acad. Sci.*, vol. 117, no. 52, pp. 33051–33060, 2020.
- [6] A. Davoudi et al., “Intelligent ICU for autonomous patient monitoring using pervasive sensing and deep learning,” *Sci. Rep.*, vol. 9, no. 1, pp. 1–13, 2019.
- [7] W. Jin et al., “Deep learning identifies synergistic drug combinations for treating COVID-19,” *Proc. Nat. Acad. Sci.*, vol. 118, no. 39, 2021, Art. no. e2105070118.
- [8] C. Xu, A. Alaa, I. Bica, B. Ershoff, M. Cannesson, and M. Schaar, “Learning matching representations for individualized organ transplantation allocation,” in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 2134–2142.
- [9] K. Stacke, G. Eilertsen, J. Unger, and C. Lundström, “Measuring domain shift for deep learning in histopathology,” *IEEE J. Biomed. Health Informat.*, vol. 25, no. 2, pp. 325–336, Feb. 2021.
- [10] A. V. Bustamante, J. Chen, L. F. Beltrán, and A. N. Ortega, “Health policy challenges posed by shifting demographics and health trends among immigrants to the United States: Study examines examine recent trends in immigrant health and health care after the great recession and the national implementation of the affordable care act,” *Health Affairs*, vol. 40, no. 7, pp. 1028–1037, 2021.
- [11] G. E. Zoeller, B. L. Drew, C. W. Schmidt, R. Peterson, and J. J. Wilson, “A paleodemographic assessment of mortality and fertility rates during the second demographic transition in rural central Indiana,” *Amer. J. Hum. Biol.*, vol. 34, no. 1, 2021, Art. no. e23571.
- [12] M. C. Odden, P. G. Coxson, A. Moran, J. M. Lightwood, L. Goldman, and K. Bibbins-Domingo, “The impact of the aging population on coronary heart disease in the United States,” *Amer. J. Med.*, vol. 124, no. 9, pp. 827–833, 2011.
- [13] Y. C. Wang, K. McPherson, T. Marsh, S. L. Gortmaker, and M. Brown, “Health and economic burden of the projected obesity trends in the USA and the UK,” *Lancet*, vol. 378, no. 9793, pp. 815–825, 2011.
- [14] J. Lei et al., “HPV vaccination and the risk of invasive cervical cancer,” *New England J. Med.*, vol. 383, no. 14, pp. 1340–1348, 2020.
- [15] S. Zhang, H. Xu, L. Zhang, and Y. Qiao, “Cervical cancer: Epidemiology, risk factors and screening,” *Chin. J. Cancer Res.*, vol. 32, no. 6, pp. 720–728, 2020.
- [16] H. Li, W. Lu, A. Wang, H. Jiang, and J. Lyu, “Changing epidemiology of chronic kidney disease as a result of type 2 diabetes mellitus from 1990 to 2017: Estimates from global burden of disease 2017,” *J. Diabetes Investigation*, vol. 12, no. 3, p. 346–356, 2021.
- [17] S. Desmet et al., “Dynamic changes in paediatric invasive pneumococcal disease after sequential switches of conjugate vaccine in Belgium: A national retrospective observational study,” *Lancet Infect. Dis.*, vol. 21, no. 1, pp. 127–136, 2021.
- [18] P. Venkatesan, “The changing demographics of COVID-19,” *Lancet Respir. Med.*, vol. 8, no. 12, 2020, Art. no. e95.

- [19] E. Nicastrì et al., "The evolution of clinical and scientific evidence-based management guidelines for Covid-19 disease: A longitudinal descriptive analysis," *SSRN Electron. J.*, 2020.
- [20] E. Mahase, "Covid-19: How many variants are there, and what do we know about them?," *BMJ*, vol. 374, 2021.
- [21] G. A. Roth et al., "Trends in patient characteristics and COVID-19 in-hospital mortality in the United States during the COVID-19 pandemic," *JAMA Netw. Open*, vol. 4, pp. e218828–e218828, 2021.
- [22] A. A. Soltan et al., "Rapid triage for COVID-19 using routine clinical data for patients attending hospital: Development and prospective validation of an artificial intelligence screening test," *Lancet Digit. Health*, vol. 3, no. 2, pp. e78–e87, 2021.
- [23] A. Youssef et al., "Development and validation of early warning score systems for COVID-19 patients," *Healthcare Technol. Lett.*, vol. 8, no. 5, pp. 105–117, 2020.
- [24] V. Losing, B. Hammer, and H. Wersing, "Incremental on-line learning: A review and comparison of state of the art algorithms," *Neurocomputing*, vol. 275, pp. 1261–1274, 2018.
- [25] Z. Chen and B. Liu, *Lifelong Machine Learning* (Synthesis Lectures on Artificial Intelligence and Machine Learning Series), 2nd ed. San Rafael, CA, USA: Morgan & Claypool, 2018.
- [26] A. Koivu et al., "Adaptive risk prediction system with incremental and transfer learning," *Comput. Biol. Med.*, vol. 138, 2021, Art. no. 104886.
- [27] N. Jackson et al., "The quality of vital signs measurements in electronic medical records varies by hospital, specialty, and patient demographics," *medRxiv*, 2022. [Online]. Available: <https://www.medrxiv.org/content/early/2022/01/21/2022.01.19.22269544>
- [28] T. T. Lee and A. S. Kesselheim, "US food and drug administration precertification pilot program for digital health software: Weighing the benefits and risks," *Ann. Intern. Med.*, vol. 168, no. 10, pp. 730–732, 2018.
- [29] L. L. Guo et al., "Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, 2022.
- [30] T. Alves, A. Laender, A. Veloso, and N. Ziviani, "Dynamic prediction of ICU mortality risk using domain adaptation," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 1328–1336.
- [31] J. Armstrong and D. Clifton, "Continual learning of longitudinal health records," 2021, *arXiv:2112.11944*.
- [32] D. Kiyasseh, T. Zhu, and D. A. Clifton, "CLOPS: Continual learning of physiological signals," 2020, *arXiv:2004.09578*.
- [33] M. Lenga, H. Schulz, and A. Saalbach, "Continual learning for domain adaptation in chest x-ray classification," in *Proc. Med. Imag. Deep Learn.*, 2020, pp. 413–423.
- [34] S. Srivastava, M. Yaqub, K. Nandakumar, Z. Ge, and D. Mahapatra, "Continual domain incremental learning for chest x-ray classification in low-resource clinical settings," in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health*. New York, NY, USA: Springer, 2021, pp. 226–238.
- [35] J. Shu et al., "Meta-weight-Net: Learning an explicit mapping for sample weighting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 1919–1930.
- [36] Y. Xu, P. Cao, Y. Kong, and Y. Wang, " $\mathcal{L}_{dmi}$ : A novel information-theoretic loss function for training deep nets robust to label noise," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, vol. 32, pp. 6225–6236.
- [37] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-learning regularization prevents memorization of noisy labels," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 20331–20342, 2020.
- [38] A. K. Menon, A. S. Rawat, S. J. Reddi, and S. Kumar, "Can gradient clipping mitigate label noise?," in *Proc. Int. Conf. Learn. Representations*, 2019.
- [39] A. Prabhu, P. H. Torr, and P. K. Dokania, "GDumb: A simple approach that questions our progress in continual learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 524–540.
- [40] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, "Normalized loss functions for deep learning with noisy labels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 6543–6553.