Máster de Metodología en las Ciencias del Comportamiento y de la Salud

Effectiveness of Continual Learning Strategies on CNN-Based Models for Glaucoma Detection

Autor: Diego Hernández Jiménez

Tutor: Luis Jáñez Escalada

Modalidad: Trabajo de investigación

Mayo de 2023

UAM
Universidad Autónoma
de Madrid

UNIVERSIDAD COMPLUTENSE
MADRID

UNED

Abstract

Glaucoma is a leading cause of irreversible blindness worldwide, and early detection is crucial for effective treatment. Convolutional neural networks (CNNs) have shown promise in detecting glaucoma from retinal images, but they fail to generalize to unseen images from different datasets. This study investigates the generalization and continual learning capabilities of CNN models in glaucoma detection. We explore the potential usefulness of Contrast Limited Adaptive Histogram Equalization (CLAHE) to improve cross-dataset generalization and Memory Aware Synapses (MAS) as a domain incremental learning approach. For that purpose, experiments are conducted on three public datasets. Results indicate that CLAHE can be beneficial to generalization in comparison with no normalization but it cannot be said that MAS performs better than naïve fine-tuning. Although not confirmed, we postulate this could be due to the presence of high label noise in fundus images, a problem also pointed out by other authors. Our study sheds light on the limitations and opportunities of CNN models in glaucoma detection and highlights the need for robust methods that can handle not only domain shift, but also label noise.

Keywords: glaucoma, continual learning, Memory Aware Synapses, CLAHE

Resumen

El glaucoma causa ceguera irreversible y la detección temprana es esencial para un tratamiento eficaz. Las redes neuronales convolucionales (CNN) tienen potencial para detectar glaucoma en imágenes de retina, pero no pueden generalizar a imágenes de otros conjuntos de datos. Este estudio investiga las capacidades de generalización y aprendizaje continuo de los modelos de CNN en la detección del glaucoma. Se prueba la potencial utilidad de la Ecualización Adaptativa de Histograma con Contraste Limitado (CLAHE) para mejorar la generalización fuera de distribución y de Sinapsis con Memoria (MAS) como un enfoque de aprendizaje incremental de dominio. Los experimentos se realizan en tres conjuntos de datos públicos. Los resultados muestran que CLAHE puede ser beneficioso para la generalización en comparación con no normalización, pero no se puede decir que MAS sea mejor que el ajuste fino ingenuo. Se postula que esto podría deberse al alto ruido de etiquetado en imágenes de fondo de ojo. Este estudio arroja luz sobre las limitaciones y oportunidades de los modelos de CNN en la detección del glaucoma y destaca la necesidad de métodos robustos que puedan manejar no solo el cambio de dominio, sino también el ruido de etiquetado.

Palabras clave: glaucoma, aprendizaje continuo, Sinapsis con Memoria, CLAHE

Contents

1.1 General theoretical background ... 4
1.2 The problem of generalization ... 5
1.3 Continual learning ... 5
1.3.1 Memory Aware Synapses ... 7
1.4 CLAHE ... 9
1.5 Related work ... 10
1.6 Proposal ... 11
2. Method ... 11
2.1 Data ... 11
2.2 Pre-processing ... 12
2.3 Experiment design ... 14
2.4 Network architecture ... 15
2.5 Learning process ... 17
2.6 Performance metrics ... 19
2.7 Statistical tests ... 19
3. Results ... 19
3.1 No normalization vs. CLAHE ... 19
3.2 Continual learning ... 20
3.2.1 Joint training ... 20
3.2.2 Naïve fine-tuning vs. MAS ... 21
4. Discussion ... 25
4.1 Limitations ... 26
4.2 Conclusions ... 27
5. References ... 29
6. Appendix A ... 33
7. Appendix B ... 36

1.1 General theoretical background

Some recent work (Zhang et al., 2021) has estimated that there more than 68 million people over 40 years old with primary angle glaucoma, the most prevalent subtype of glaucoma, a group of eye diseases characterized by elevated intraocular pressure (Casson et al., 2012). Glaucoma symptoms start slowly and may not be detected in early stages, but it progresses and can lead to significant visual loss. In fact, it is the first cause of irreversible blindness in the world.

As early diagnosis is the key factor to slow down the progression of the disease, there is a great need of accurate cost-effective screening methods to detect it as soon as possible. In addition, those methods must have high sensitivity and specificity regardless of the context in which they are applied. Given that there is not a unique way of diagnosing glaucoma, clinicians must rely on different sources of information to make a decision. Common medical tests are based on examination of visual field (VF), optical coherence tomography (OCT) images and fundus photos (Chen et al., 2023). In contrast with VF tests and OCT, color fundus images can be acquired using relatively inexpensive portable fundus cameras and the photographs or retinographies can be utilized not only for glaucoma diagnosis, but also for the diagnosis of other eye diseases (Mirzania et al., 2021).

Visual inspection and interpretation of fundus images by medical experts is very costly and time consuming and, for that reason, many machine learning based detection methods have been developed. The key advantage is that they automate the process of diagnosis and classification from images. Support Vector Machines (SVM) and Random Forests (RF) are some of the more traditional machine learning techniques employed (e.g. Dey & Bandyopadhyay, 2016), but deep learning algorithms like Convolutional Neural Networks (CNN) systematically outperform the rest of methods (e.g. Shibata et al., 2018). This is in part because, contrary to SVMs and RFs, convolutional networks are designed to effectively leverage spatial and configurational information by utilizing images as their input (Shen et al., 2017). Another advantage of CNNs is that they not only discover the mapping from the features to the output, but they also automatically extract the features themselves and this learned representations often result in better performance than hand-designed features (Goodfellow et al., 2016 p.4). This is extremely useful in complex visual tasks such as glaucoma diagnosis because it is very difficult to manually identify and implement the visual features observed by professional ophthalmologists when examining an image.

1.2 The problem of generalization

CNNs have been used extensively in the field of glaucoma detection; different architectures have been tested and multiple situations have been explored. However, the topic of generalization has received very little attention. Most of the studies assess generalization in terms of performance on an unseen subset of images of the *same* dataset that was used for training. Still, that performance might not be a good indicator of the true generalizability. To the best of our knowledge, Díaz-Pinto et al. (2019) were the ones to first report model performance on unseen images from different datasets than the ones used for training and they showed that accuracy was reduced by approximately 15% when the test set came from a different dataset. Fumero et al. (2020) also noticed a slight reduction in performance in a similar situation in which test images were taken at different hospitals with different cameras, although the scenario was not identical, as retinographies were annotated by the same experts.

There are several possible reasons for the poor generalizability. First, CNNs are trained solely on the characteristics that exist in the dataset used for training. Consequently, deep learning algorithms may exhibit suboptimal performance when applied to images captured using dissimilar cameras, or images of eyes affected by concurrent retinal and optic nerve conditions that were not included in the training dataset (Mirzania et al., 2021). Secondly, because the ground truth of the fundus images is based on human judgments, which do not tend to have high inter-observer agreement (Mirzania et al., 2021), some decrease in performance could be expected when classifying images from a dataset collected by different researchers. A similar argument is posed by Díaz-Pinto et al. (2019), who state that fundus images are usually labeled in two fundamentally different ways: diagnosis can be done using only the image itself or using the retinography and clinical data from the patient. Because the former method relies just on visual information, it increases the label noise, making the generalization more difficult.

1.3 Continual learning

To develop a robust CNN classifier under these circumstances means to build a model that can effectively perform on images from both old and new datasets, even when training only on data from the most recent dataset (Jung et al., 2018). That is because we are in a non-stationary environment where new datasets appear from time to time and must be learned sequentially (Hadsell et al., 2020). Even though the classification task is always the same, it is clear for the reasons explained above that images from different datasets are not identically distributed, they constitute different domains. To put it more formally, we have d domains $\mathcal{D}_1, \dots, \mathcal{D}_d$, containing N feature-label pairs: $\mathcal{D}_t = \left\{\left(x_i^{(t)}, y_i^{(t)}\right): x_i^{(t)} \in \mathcal{X}, y_i^{(t)} \in \mathcal{Y}\right\}_i^N$ where $\mathcal{X}$ is the set of features and $\mathcal{Y}$ the set of class labels. The output space for class labels is shared for all domains, meaning that $\mathcal{Y}_t = \mathcal{Y}_{t'}$ with $t \neq t'$. Also, for each domain we have a joint probability distribution $p_t(\mathcal{X}, \mathcal{Y})$ that is different from others: $p_t(\mathcal{X}, \mathcal{Y}) \neq p_{t'}(\mathcal{X}, \mathcal{Y})$, for $t \neq t'$.

Given all these characteristics, it is sensible to frame the problem from the perspective of continual learning (Hadsell et al., 2020) or, more precisely, domain incremental learning (van de Ven et al., 2022). In this work, both terms will be used interchangeably. Note that unlike the related concept of domain adaptation, which seeks to transfer knowledge from a prior task to a new task and only the performance on the new task is considered, the setting of continual learning aims to maintain performance on prior domains while also achieving reasonable performance on the new domain (Hsu et al., 2018)

This theoretical framework provides a good foundation to design CNN models that can learn sequentially, but any solution proposed from the continual learning perspective must also satisfy some basic desiderata (Hadsell et al., 2020), which we present next:

During training, the model should not have access to previously seen domains. This is especially necessary in the context of glaucoma classification and medical images classification in general because access to data is usually restricted and datasets might not always be available. Also, if previous data can be stored and reused, there is no need for continual learning, we could just retrain the model again from scratch with all the data. In fact, this joint training approach is usually used as an upper bound for performance when assessing continual learning strategies.

Model capacity should be constant or minimally increased. That implies that the naïve solution involving training multiple independent models, one for each domain, is not a viable solution since there is no generalization at all.

The approach should minimize interference and catastrophic forgetting when training on new domains. Catastrophic forgetting is a well-studied phenomenon that is observed when a model must learn a new task, or in our case, a new domain. While the performance on the new domain is improved, the performance on previously learned domains is reduced drastically. (McCloskey & Cohen, 1989; French, 1999)

During learning and inference, the model should not be aware of which domain the input data had come from. In our domain incremental learning scenario, the structure of the problem is always the same, the class labels are static, but the source dataset of the images varies. The designed model must not need to consider this information to make a good prediction, it must rely only in the visual information provided by the images. This restriction creates domain-agnostic models.

1.3.1 Memory Aware Synapses

One method that satisfies all the conditions is the strategy of Memory Aware Synapses (MAS) (Aljundi et al., 2018). It was proposed as a class incremental learning algorithm, but it can also be applied in a domain incremental learning context. It belongs to the regularization-based methods family where the focus is put on the loss function, which is modified to balance the ability to retain knowledge from previous domains (stability) and the ability to learn features from new domains (plasticity) (De Lange et al., 2022).

MAS is inspired by other approaches like Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) and works similarly. The main idea is to penalize changes in parameter values that are deemed important for already learned tasks with the aim of mitigating the effect of catastrophic forgetting. Given a loss function L for a task, like cross-entropy, the MAS loss is computed as in (1):

$$L_{MAS}\left(f\left(\mathbf{x}_i; \boldsymbol{\theta}^{(t)}\right), y_i\right) = L\left(f\left(\mathbf{x}_i; \boldsymbol{\theta}^{(t)}\right), y_i\right) + \lambda \sum_{j}^{k} \Omega_j^{(t-1)} \left(\theta_j{}^{(t)} - \theta_j{}^{(t-1)}\right)^2 \qquad (1)$$

Where f represents the learned function, that is, the neural network, which takes an image $\mathbf{x}_i$ as input. The image is a multidimensional array or tensor $\mathbf{x}_i \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels (which is three for RGB images), H is the height of the image and W is the width. The network is parametrized by the set of parameters $\boldsymbol{\theta}^{(t)}$ of domain t (in vector form for convenience). Thus, $f\left(\mathbf{x}_i; \boldsymbol{\theta}^{(t)}\right)$ is the prediction of the model for the input $\mathbf{x}_i$ of domain t. The true label for $\mathbf{x}_i$ is y_i, having $y_i \in \{0,1\}$ in the binary classification setting. The hyperparameter λ controls the amount of regularization and, finally, $\boldsymbol{\Omega}$ encodes the parameter importance. The dimension of this vector equals the number of trainable parameters, k. The importance term reflects the sensitivity of the learned function to changes in parameter values and is based on a known result from calculus. In the simple case where we only have one parameter θ, if we take δ to be a small perturbation added to θ, then the difference in the output of f can be approximated by the derivative of f multiplied by the perturbation quantity:

$$f(\mathbf{x}_i;\theta+\delta)-f(\mathbf{x}_i;\theta)\approx\frac{df(\mathbf{x}_i;\theta)}{d\theta}\delta \tag{2}$$

Assuming a constant perturbation, we can use the derivative as a measure of sensitivity. In the multivariate setting with N data points (belonging to the training set, for example), importance for parameter j is then estimated as the average sensitivity over the given set of exemplars:

$$\Omega_j=\frac{1}{N}\sum_{i}^{N}\left\|\frac{\partial f(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_j}\right\| \tag{3}$$

The Euclidean norm is used to avoid negative scores. When the output of f is multi-dimensional, the authors propose to take the partial derivatives of the squared ℓ_2 norm of f, but in a binary case with only one output unit this is just equivalent to the derivative of f. In any case, the argument of the norm is then no longer a vector, so parameter importance can be rewritten:

$$\Omega_j=\frac{1}{N}\sum_{i}^{N}\left|\frac{\partial f(\mathbf{x}_i;\boldsymbol{\theta})}{\partial\theta_j}\right| \tag{4}$$

A cumulative moving average of the importance values is kept and updated every time a new task is learned (Aljundi et al., 2019). The main advantage of the MAS approach over EWC is that importance weights can be estimated in an unsupervised manner because there is no need of class labels.

Additionally, although it was not made in the original paper, it can be beneficial to normalize importance scores. As noted by Schwarz et al. (2018), this practice allows the model to update the values of $\boldsymbol{\Omega}$ based on the relative importance of network parameters, treating each task equally. Through the rest of the paper, any mention to the importance term will be referring to the normalized version.

1.4 CLAHE

The MAS approach focuses on modifying training to reduce catastrophic forgetting, but other complementary and more general techniques can also be used to deal with the domain shift when trying to generalize performance to out-of-distribution samples. As it was said, diversity in cameras and patients can increase noise in images and prevent the model from learning domain-invariant features. Image normalization techniques can greatly help in the task of making the input more homogeneous. One of those methods is Contrast Limited Adaptive Histogram Equalization (CLAHE) (Zuiderveld, 1994). Technically, CLAHE is not a normalization method but an image enhancement technique, and it has not been used in the context of incremental learning, but it can be useful, nonetheless. The key algorithm behind CLAHE is histogram equalization, which works by

> (…) assigning brightness values of pixels based on the image histogram. Individual pixels retain their brightness order (that is, they remain brighter or darker than other pixels, except in cases where pixels with several initial values are assigned the same equalized value) but the values are shifted, so that as an average, an equal number of pixels have each possible brightness value. (Russ, 2011, pp. 275)

Mathematically, the image histogram counts the number of pixels that have intensity v and the cumulative histogram of v, $h(v)$, counts the number of pixels with brightness equal to or less than v. Now, assuming the image has pixel intensities in the range $[a, b]$, histogram equalization involves applying min-max scaling (Han et al., 2012) to the cumulative histogram values:

$$v' = round\left(a + \frac{h(v) - h_{min}}{H \cdot W - h_{min}}(b - a)\right) \tag{5}$$

Where H stands for height and W for width and their product gives the total number of pixels in the image, which is always h_{max}, the maximum value of the cumulative histogram. Similarly, h_{min} is the minimum value of the cumulative histogram. Alternatively, we can use the cumulative distribution function, which results from dividing the cumulative histogram values by the total number of pixels:

$$cdf(v) = \frac{h(v)}{H \cdot W} \tag{6}$$

In that case, and further assuming that pixel values lie in the range $[0,255]$, we arrive to the formula more commonly used (Gonzalez & Woods, 2018):

$$v' = round(cdf(v) \cdot 255) \tag{7}$$

Histogram equalization can enhance image contrast, but it can also result in an unrealistic and unnatural appearance. CLAHE overcomes this limitation by applying histogram equalization locally in the image, in patches or neighborhoods of pixels (Zuiderveld, 1994; Russ, 2011). This adaptive approach ensures that the contrast enhancement is localized and does not affect the overall appearance of the image. Moreover, CLAHE sets a maximum threshold value for the amount of contrast enhancement that can be applied to each region to prevent over-amplification of noise or other unwanted artifacts (Zuiderveld, 1994).

By applying CLAHE to fundus images that may have, for instance, different lighting conditions, their contrast and visibility can be improved, making it easier for a model to extract meaningful representations from the images while preserving its local features. Therefore, in principle, with CLAHE we can help ensure that CNN models are better equipped to handle a wide range of lighting conditions and improve their out-of-distribution generalization.

1.5 Related work

Up to date, only Díaz-Pinto et al. (2019) have actively tried to solve the poor generalization problem in glaucoma prediction, but they have not explicitly approached the problem from a continual learning perspective. In their work, they propose to “retrain the CNNs when trying to classify images from databases different to the ones used for training” (p.16), which is just a form of sequential fine-tuning, and can be considered as a naïve continual learning solution because it does not actively protect against catastrophic forgetting. Nonetheless, this proposal was not empirically tested in the paper, so its potential efficacy has not been evaluated yet.

On the other hand, there is a wealth of studies that have experimented with MAS and other continual learning strategies (see De Lange et al. (2022) and Masana et al. (2022) for recent surveys), but most of them are concerned with the class and task incremental learning scenarios, not with domain incremental learning (van de Ven et al., 2022), and only one of them have applied the algorithms to medical images (Derakhshani et al., 2022). Unfortunately, fundus images for glaucoma diagnosis were not used in their experiments.

Finally, because CLAHE has conventionally been treated solely as a pre-processing step (Gupta et al., 2021; Shoukat et al., 2021), its potential to enhance the robustness of classifiers and improve cross-dataset generalization has largely remained unexplored.

1.6 Proposal

Despite the lack of studies tackling this specific problem in this context, there is enough theoretical background to postulate the hypothesis that a continual learning strategy such as MAS can help learn new domains while reducing the effect of catastrophic forgetting of old domains. More concretely, we argue that, in comparison with a baseline like naïve fine-tuning, MAS would perform better, in the sense that, after having learned t domains, one at a time, MAS would have better performance on those t domains than fine-tuning. We also suggest that an image enhancement technique like CLAHE can act as an input normalization method and reduce domain variability and thus benefit generalization on unseen domains. Still, both hypotheses will be translated into two-tailed significance tests because there is not enough prior evidence to make the strong specific directional predictions needed to justify the use of one-tailed tests.

The main contributions of this work are two-fold. For the first time in the context of glaucoma research, the problem of lack of generalization has been fully addressed and the continual learning framework has been proposed as a useful way of conceptualizing the problem. Three approaches have been evaluated: a general method (CLAHE) and two more specialized strategies (naïve fine-tuning and MAS).

2. Method

All data processing and analysis tasks were implemented using Python programming language version 3.10.11 (Van Rossum & Drake, 2009) in a Google Colab environment with GPU enabled. Pytorch version 2.0.0 (and CUDA 11.8) (Paszke et al., 2019) was the framework of choice for deep learning modeling. A snippet of the code is provided in the Appendix A. The rest of the code will be made publicly available in the future.

2.1 Data

Experiments were conducted on three publicly available datasets: ACRIMA (Díaz-Pinto et al., 2019), RIM-ONE-DL (Fumero et al., 2020) and PAPILA (Kovalyk et al., 2022).

ACRIMA: It is composed of 705 fundus images (396 of them from glaucomatous patients). The sample comes from Spanish population, but no further data on demographics is given. Images were labeled by two experts with 8 years of experience in the field of glaucoma, with each one being responsible of a subset of the retinographies.

RIM-ONE-DL: It is composed of 485 fundus images (172 of them from glaucomatous patients). Retinographies were obtained from patients in three different Spanish hospitals and with different cameras. No more data about the composition of the sample is given. Images were labeled jointly by two experts in the field and a third expert, with 20 years of experience, decided the diagnosis in case of disagreement.

PAPILA: The original dataset is composed of 488 fundus images, but images correspond to three types of patients: healthy, glaucomatous and suspect of having glaucoma. Due to the uncertainty associated with the labeling of suspect patients, we are not considering this category. As a result, the dataset finally used contains 420 images (67 of them from glaucomatous patients). All the images were collected in a single Spanish hospital and always with the same camera. Further details of the sampling process and composition of the dataset are provided by the authors in the original article. Two experts annotated the images, each one being responsible of a subset of the retinographies. In contrast with ACRIMA and RIM-ONE-DL annotators, who based their diagnosis only on visual inspection, PAPILA experts had access to other measurements and tests apart from the retinographies to make a decision.

2.2 Pre-processing

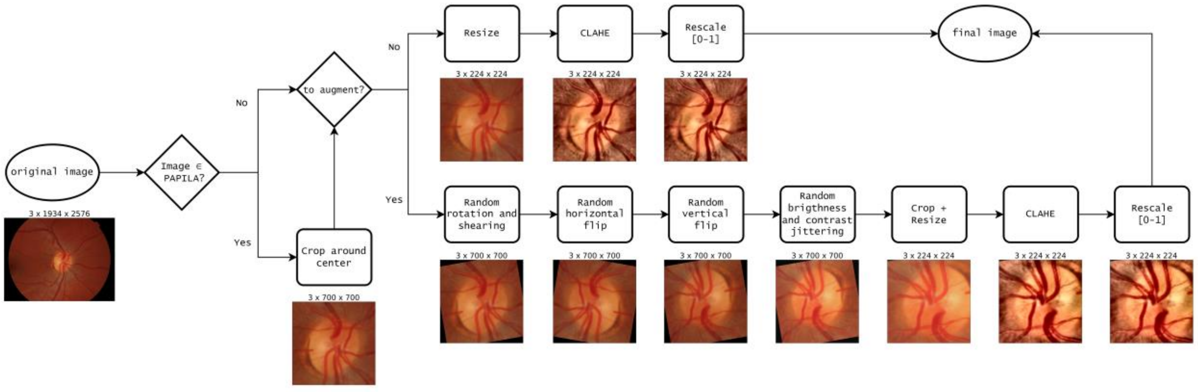
The fundus images were processed before being fed to the neural network (see Figure 1 for a visualization of the pre-processing workflow). Images coming from PAPILA dataset had to be manually cropped around the optic disc. To do this, optic disc centers coordinates provided by Kovalyk et al. (2022) were taken as centroids of their respective images. Then, a square crop was made around those centers, with every side of the image at a distance of 350 pixels of the center.

Retinographies were first resized to a common resolution of 224×224. The images were then transformed using CLAHE algorithm. This method is not well suited for RGB images (Russ, 2011) so the original retinographies were converted to L*a*b color space. The CLAHE algorithm was then applied to the L channel. The chosen size of the neighborhood was 5×5 and the limit of the contrast enhancement was set to 5. Finally, the images were converted back to RGB. After the process of normalization with CLAHE, the images were rescaled to make every pixel to be in the range [0,1].

To artificially generate more samples, data augmentation techniques were employed. The sequence of transformations used is depicted in Figure 1. An affine transformation keeping the center invariant is applied to the original image. This involves the rotation of d degrees, where d is uniformly selected from the range $[-10, 10]$, and shearing of s degrees, where s is uniformly selected from the range $[-0.2, 0.2]$. The image is then vertically flipped with probability of 0.5 or horizontally flipped with probability of 0.5. Brightness and contrast were also increased by a factor of b and c respectively, where b is uniformly selected from the range $[0.7\ 1.5]$. and c is uniformly selected from the range $[0.5\ 1.5]$. The last transformation consisted of cropping a random portion of the image with area 224^2 and then resizing it to be 224×224. This technique is similar to zooming on a random section of the retinography. CLAHE and rescaling were ultimately applied to the resulting augmented images. It is important to note that because online data augmentation was performed, meaning that artificial samples were created during training, the augmented images were different every epoch of training.

Figure 1

Pre-processing workflow diagram



Note. Flow diagram representing the sequential steps followed during pre-processing with an example of their effect on an image.

2.3 Experiment design

Three conditions were compared in the experiment: naïve fine-tuning, MAS and joint training approaches, with the latter acting as an upper bound reference. One model was trained sequentially in three steps using one dataset at a time and always with the same order: ACRIMA, RIM-ONE-DL, PAPILA. Performance was recorded on each of the steps. Performance at step 1, when only the first domain had been seen, was taken as lower bound reference or baseline of cross-dataset generalization. Step 1 scores were also used to evaluate the effect of CLAHE.

The experiment was performed in two stages. In the first stage, manual hyperparameter tuning was carried out on each of the datasets as in a non-continual learning scenario using repeated holdout validation (Raschka, 2018). Decisions about hyperparameters were only based on results obtained on the dataset being used and not on results of the others, which could constitute a form of data leakage. Following conventional practice, data was randomly split into two sets: 75% of the sample was used as training data, and the remaining 25% of the sample was used as validation set. There was no test set due to the limited sample size. Data augmentation was used to double the number of observations in the training set. In the specific case of RIM-ONE-DL and PAPILA, where there is a severe imbalance of classes, oversampling of the minority class (glaucomatous cases) was used to rebalance the distribution of normal and glaucomatous samples. With the purpose of increasing heterogeneity among the oversampled observations, data augmentation preprocessing was applied to them. The resulting sample sizes of training and validation sets can be seen in Table 1.

In the second stage, which consisted of performance estimation, the values for the learning rates, decay schedule, batch size… were kept fixed. In order to get unbiased performance estimates, 4-fold cross validation was employed. This method avoids possible selection biases derived from the limited sample size of the datasets because it guarantees that every image is used for training and for validation (Raschka, 2018). The k-fold cross validation achieves that by partitioning the data into k disjoint sets. Training is done with $k-1$ of those sets, and the remaining part is used as validation set to get an estimate of model performance. This process is repeated k times, and each time a different subset is used for validation. The final cross validation performance estimate is the arithmetic average of the k computed metrics (Raschka, 2018). In non-deep learning settings, k is typically set to 5 or 10.

In this context, however, it would entail a high computational cost, so k was chosen to be 4, which was close to the conventional $k = 5$ but also ensured that the data split ratio was 75:25, as in the previous stage. Data augmentation and oversampling was done as before.

In the case of joint training, the only difference resided in the fact that the training sets of the three datasets were concatenated as one. The same was true for the validation sets. Cross validation was performed only once.

Table 1

Sample sizes for training and validation sets

Dataset	**Glaucoma**			**Normal**		**Total**	
	orig.	aug.	oversamp.	orig.	aug.	N	% glaucoma
ACRIMA	297 | 99	297 | 0	0 | 0	231 | 78	231 | 0	1056 | 177	56 | 56
RIM-ONE-DL	129 | 43	129 | 0	258 | 0	234 | 79	234 | 0	984 | 122	52 | 35
PAPILA	65 | 22	65 | 0	325 | 0	250 | 83	250 | 0	955 | 105	48 | 21

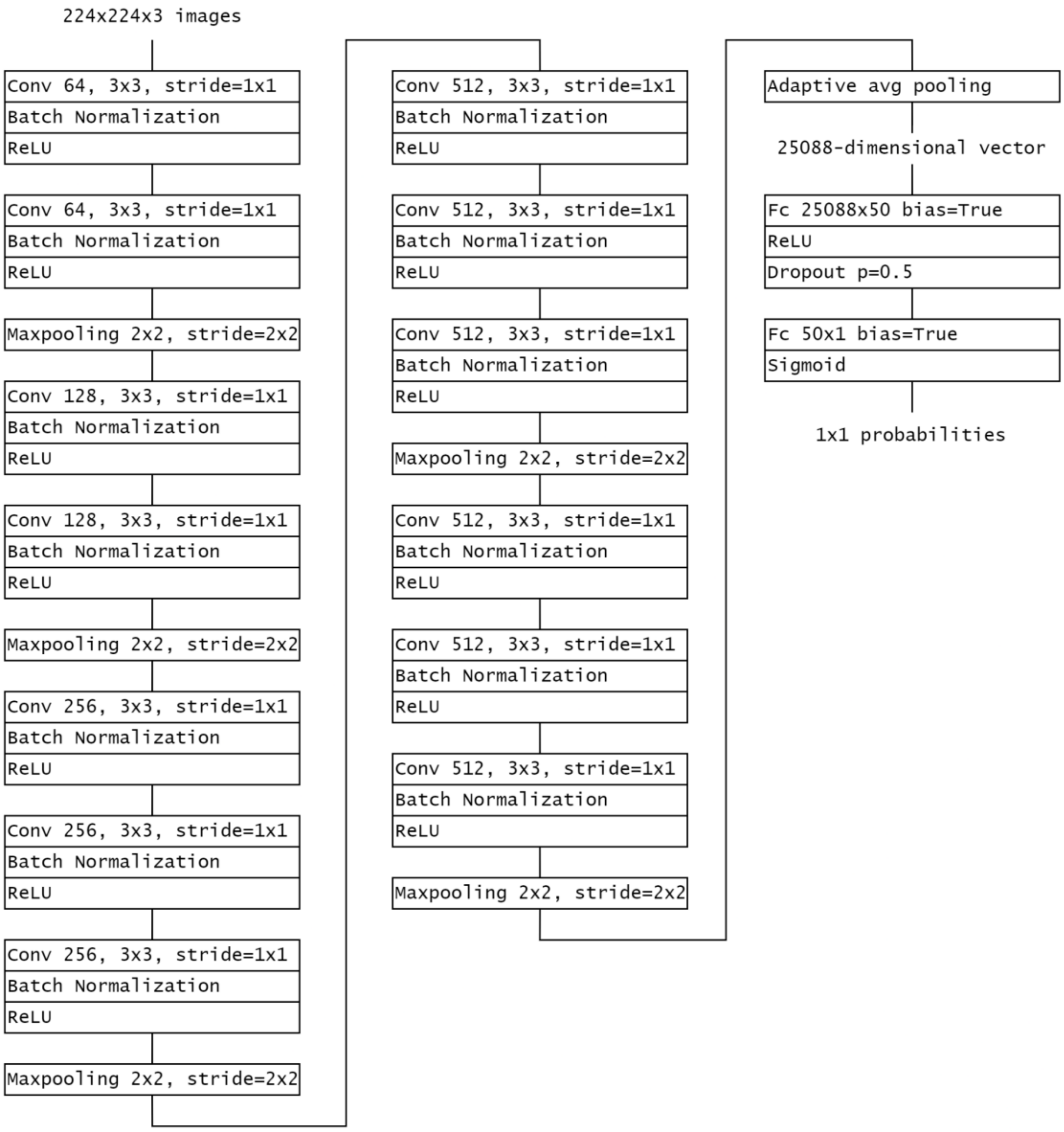
Note. Orig stands for original, *aug* for augmented and *oversamp* for oversampled. In each cell, the value from the left is the number of observations in the training set for a given dataset and condition, whereas the right value corresponds to the number of observations in the validation set. The numbers from the last column do not represent number of cases but percentage of images labeled as glaucoma.

2.4 Network architecture

In this work, the CNN architecture of choice was VGG16 (Simonyan & Zisserman, 2014). More specifically, a modified version of the D configuration with batch normalization layers (Ioffe & Szegedy, 2015) between convolutional layers was used. Despite the simplicity of the architecture, it has shown to reach state-of-the-art or near state-of-the-art performance in glaucoma classification (Díaz-Pinto et al., 2019; Gómez-Valverde et al., 2019; Batista et al., 2020). In addition, the classifier head of the model, a perceptron with two hidden layers, was replaced by perceptron with one single hidden layer of 50 neurons with Rectified Linear Unit (ReLU) activation function and dropout (Goodfellow et al., 2016). Because the target variable was binary, output layer had one unit and the sigmoid function was used to get probability estimates. The change in the classifier was empirically motivated by the fact that it was observed that this simpler version was able to perform equally as well as the original unmodified network with a reduction of over 100 million parameters. Figure 2 shows the architecture in more detail.

Figure 2

Modified VGG-16 network architecture.



Note. The first two blocks, starting from the left, act as a fixed feature extractor while the two fully connected layers from the rightmost part of the image are the only trainable layers.

2.5 Learning process

The compared domain incremental learning strategies are forms of transfer learning. The process begins with the base VGG16 network, with weights that have been optimized for ImageNet (Russakovsky et al., 2015) classification (Pytorch, 2023). We adopt a *freezing layers* approach (Iman et al., 2023) where all layers except the last two were “frozen”, meaning that their parameter values were fixed (setting the learning rate associated with them to 0) so they would not update during the training phase. Because the last layers correspond to the classifier and there are not convolutional layers, we are in practice using the pretrained network as a fixed feature extractor (Murphy, 2022, pp. 625-626). The reason for this approach is the well-known fact that CNN first layers extract more general features, but last-layer features are more specific and dependent on the task (Yosinski et al., 2014). Also, by making use of a model pre-trained in a similar task (image classification) and just tuning some of the parameters of the network, the training cost is reduced while improving the generalizability (Yosinski et al., 2014).

The weights and biases from trainable fully connected layers were randomly initialized and were not kept fixed when training the first dataset. However, when learning the subsequent datasets, weights were initialized with the parameters obtained after training with the previous domain. The MAS approach works in the same way. The only difference lies in the loss function that is optimized. It includes a regularization term that penalizes changes in parameters that have been estimated to be important for previous tasks.

For cross-validation, mini-batch gradient descent with momentum (Ruder, 2017) was the optimization algorithm of choice. Batch size was 32 and momentum 0.9. The initial learning rate was set to 0.005 or 0.001, depending on the dataset (see Table 2 for details), and it was decayed by a factor of 0.5 every 3 or 5 epochs, depending on the dataset. Model was trained for 10 or 20 epochs. The number of training epochs was selected based on early stopping results during the hyper-parameter selection phase. The optimization algorithm was stopped if an improvement of 0.001 in accuracy over the previous best value was not observed after 5 consecutive epochs. The objective function to minimize was binary cross entropy (BCE), which, for a given data point, is generally defined as:

$$L_{BCE}(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i) = -[\ln f(\mathbf{x}_i;\boldsymbol{\theta}) + (1-y_i)(1- \ln f(\mathbf{x}_i;\boldsymbol{\theta}))] \tag{8}$$

With N training data points, the average cross-entropy is what is minimized:

$$BCE = \frac{1}{N}\sum_{i}^{N} L_{BCE}(f(\mathbf{x}_i;\boldsymbol{\theta}), y_i) \tag{9}$$

Note that if the diagnosis y is considered to be a Bernoulli variable and $f(\mathbf{x}_i;\boldsymbol{\theta})$ is seen as the probability of glaucoma $p(y_i = 1|\mathbf{x}_i;\boldsymbol{\theta})$, then $-N \cdot BCE$ is the log-likelihood of the model and minimization of cross-entropy can thus be seen as maximum likelihood estimation. In the case of MAS, the loss function also included an additional term, as explained previously. The value of λ, which controls the degree of regularization, was chosen loosely following the heuristic rule given in Aljundi et al. (2018) (λ equals the maximum value that achieves satisfactory performance on the new domain). It was set to 1 when training with the RIM-ONE-DL dataset and when training with PAPILA.

Table 2

Hyper-parameter settings

Condition	Step	Initial learning rate	Learning rate decay schedule	Training epochs	λ
Continual learning approaches (naïve fine-tuning and MAS)	1	0.005	Reduce by 50% every 3 epochs	10	/
	2	0.001	Reduce by 50% every 5 epochs	20	1
	3	0.001	Reduce by 50% every 5 epochs	20	1
Joint training	/	0.005	Reduce by 50% every 3 epochs	10	/

Note. A slash symbol in the cell is used when the hyper-parameter is not used.

2.6 Performance metrics

The classification performance of the model was assessed on the validation data using two metrics: balanced accuracy and the Area Under the Curve (AUC).

Balanced accuracy is defined as the arithmetic mean of sensitivity and specificity. It is a useful metric in the context of imbalanced classes because it gives an equal weight to the proportion of correctly classified positive examples and the proportion of correctly classified negative examples. The score ranges from 0 to 1. It is commonly expressed as

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2} \tag{10}$$

This was the metric used as reference and for model comparison.

The AUC score represents the area under the ROC curve, which is a piecewise linear curve that results from plotting the true positive rate of the model against the false positive rate at different decision thresholds. The AUC statistic ranges from 0 to 1. This metric was computed to facilitate comparison with other works.

2.7 Statistical tests

Two-tailed significance tests with $\alpha = 0.05$ were used to compare the cross-validation results of the two continual learning approaches and normalization strategies (no normalization or CLAHE). Specifically, the corrected k-fold cross validation test, as described in Bouckaert & Frank (2004) was employed. This is just a corrected version of the t-test developed by Nadeau & Bengio (1999) to account for the fact that the overlapping of training sets in the k folds leads to underestimation of the performance estimates variance, which causes an increase in type I error. It should be noted, however, that there are no optimal solutions as no test allows us to estimate variance of cross-validation estimates unbiasedly (Bengio & Grandvalet, 2004).

3. Results

3.1 No normalization vs. CLAHE

The effect of CLAHE on generalization is presented in Table 3. The first thing to note is that when using CLAHE, balanced accuracy on the validation set from ACRIMA is reduced one point compared to when CLAHE is not used, but this reduction is non-significant ($t = -1.56$,

$p = 0.217$). However, performance on different datasets is higher when CLAHE is used, implying a better cross-dataset generalization. In the case of RIM-ONE-DL, the difference of 3 points favoring CLAHE is not statistically significant ($t = 1.232,\ p = 0.306$), but for PAPILA there is a statistically significant difference of almost 7 points ($t = 6.988,\ p = 0.006$). The cost of not using CLAHE is going from a balanced accuracy of 0.64 to a score of 0.53.

Table 3

Results for the compared normalization strategies after training with ACRIMA

Metric	Normalization	Dataset		
		ACRIMA	RIM-ONE-DL	PAPILA
Balanced accuracy	None	**0.962**(0.015)	0.515(0.014)	0.531(0.039)
	CLAHE	0.952(0.024)	**0.543**(0.038)	**0.645**(0.048)**
AUC	None	**0.996**(0.003)	0.65(0.065)	0.656(0.068)
	CLAHE	0.991(0.006)	**0.728**(0.044)	**0.669**(0.04)

Note. Scores are averaged over the 4 folds and the standard deviation is written in parenthesis. For each dataset and metric, the highest value of the two compared approaches is highlighted in bold font. Statistically significant differences of balanced accuracy scores are indicated with one asterisk ($p < 0.05$), two ($p < 0.01$) or three ($p < 0.001$) next to the highest score.

3.2 Continual learning

3.2.1 Joint training

Cross validation results from the joint training condition are presented next. In terms of balanced accuracy, score of 0.922 ($SD = 0.022$) was obtained for ACRIMA, 0.864 ($SD = 0.05$) for RIM-ONE-DL and 0.745 ($SD = 0.093$) for PAPILA. They represent the upper bound for performance because information from all domains is available at the moment of training. The AUC scores were higher in all cases: $M = 0.979$, $SD = 0.008$ for ACRIMA, $M = 0.924$, $SD = 0.038$ for RIM-ONE-DL and $M = 0.861$, $SD = 0.046$ for PAPILA, indicating an optimistic bias, which could be due to the class imbalance. As expected, AUC values are equal or better than those found when training with each dataset independently (Díaz-Pinto et al., 2019; Fumero et al., 2020; Kovalyk et al., 2022).

3.2.2 Naïve fine-tuning vs. MAS

Performance scores of the fine-tuning and MAS approaches are summarized in Table 4. Balanced accuracy values are from the last step of the learning sequence, that is, scores obtained in the three validation datasets once the model had been trained on the three domains. As expected, joint training scores are far higher than scores obtained in a continual learning scenario, but the differences between continual learning strategies are not obvious. There is a slight advantage of the naïve fine-tuning approach over MAS in RIM-ONE-DL and PAPILA, but the differences are not statistically significant ($t = -0.487,\ p = 0.66$ for RIM-ONE-DL, $t = -2.24,\ p = 0.11$ for PAPILA).

The forgetting effect can be visualized in Figure 3 and Figure 4. As it can be seen, there is a severe drop in performance on ACRIMA images when the model is naively finetuned and when MAS is used (more than ten points, regardless of the approach). The magnitude of forgetting is a bit lower in the MAS condition, but at the expense of reducing performance in the RIM-ONE-DL dataset. At step 3, the performance on ACRIMA is further reduced, but this time only by approximately 2 points in both conditions. For RIM-ONE-DL at step 3, little or no forgetting is observed on average. PAPILA domain is not seen until step 3 but the performance on that dataset gradually increases over time.

Table 4

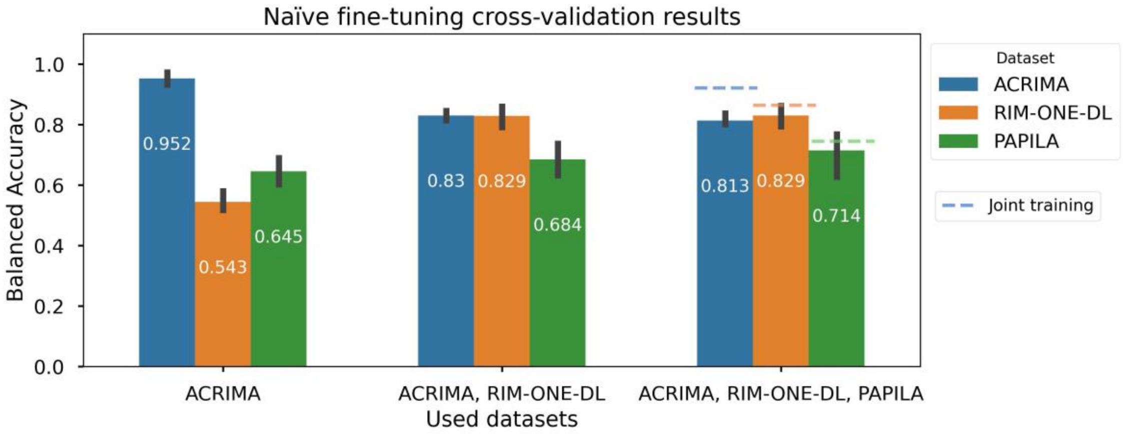
Results for the compared continual learning strategies after completing all training steps

Metric	**Strategy**	**Dataset**		
		ACRIMA	RIM-ONE-DL	PAPILA
Balanced accuracy	Naïve fine-tuning	0.813(0.024)	**0.829**(0.039)	**0.714**(0.088)
	MAS	**0.819**(0.028)	0.822(0.046)	0.689(0.087)
AUC	Naïve fine-tuning	0.89(0.012)	**0.907**(0.035)	**0.853**(0.048)
	MAS	**0.891**(0.016)	0.904(0.04)	0.847(0.039)

Note. Scores are averaged over the 4 folds and the standard deviation is written in parenthesis. For each dataset and metric, the highest value of the two compared approaches is highlighted in bold font. Statistically significant differences of balanced accuracy scores are indicated with one asterisk ($p < 0.05$), two ($p < 0.01$) or three ($p < 0.001$) next to the highest score.

Figure 3

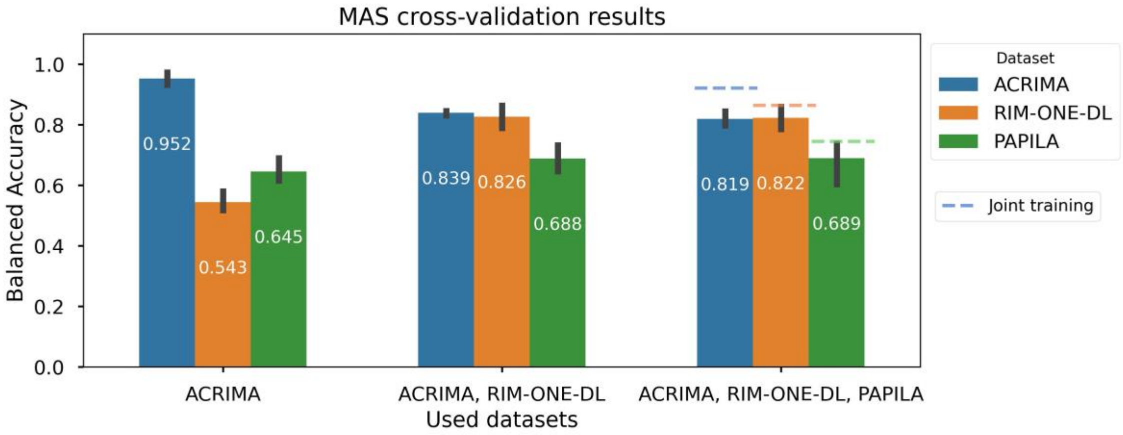
Average balanced accuracy in the three datasets at every step for naïve fine-tuning



Note. Bars represent mean balanced accuracy for each domain in the naïve fine-tuning condition. Black vertical bars indicate the 95% confidence interval for balanced accuracy scores. Dashed horizontal lines at the top of step 3, when all datasets have been used.

Figure 4

Average balanced accuracy in the three datasets at every step for MAS

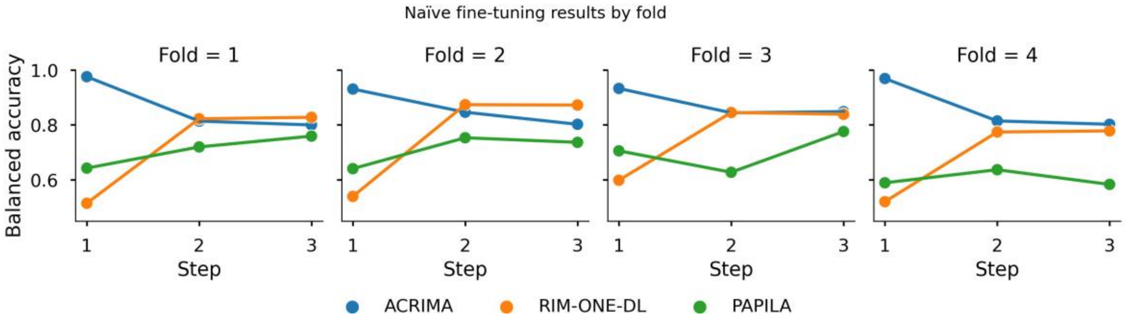


Note. Bars represent mean balanced accuracy for each domain in the MAS condition. Black vertical bars indicate the 95% confidence interval for balanced accuracy scores. Dashed horizontal lines at the top of step 3, when all datasets have been used.

Regardless of the approach, there is a great amount of variance in performance between folds, which makes the estimated scores a bit unstable. This is more noticeable in the case of PAPILA scores in the last step of training. In terms of standard deviation, the variability between folds is around 0.08. The source of this variability comes in part from fold 4, where PAPILA scores are systematically lower at every step as can be seen in Figure 5 and Figure 6.

Figure 5

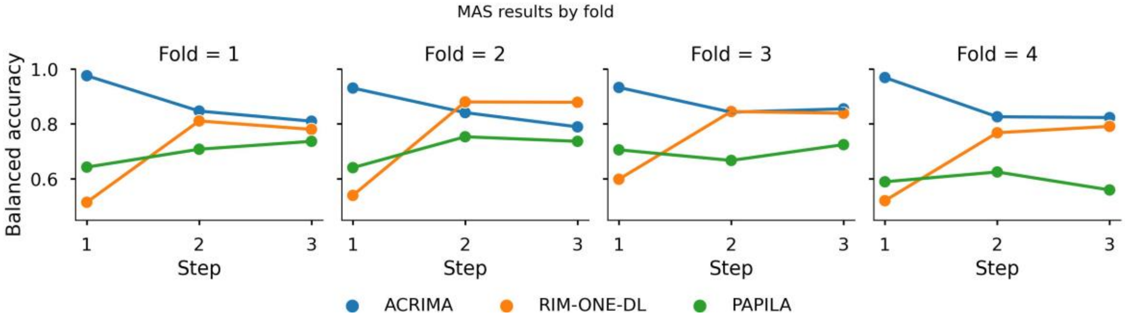
Naïve fine-tuning balanced accuracy in the three datasets at every step grouped by fold



Note. Steps on the x-axes correspond to the moments where the model has been trained on ACRIMA (step 1), ACRIMA and RIM-ONE-DL (step 2) and ACRIMA, RIM-ONE-DL and PAPILA (step 3). Notice also that the y-axis ranges from 0.5 to 1, instead of 0 to 1. This is to better visualize changes in score values.

Figure 6

MAS balanced accuracy in the three datasets at every step grouped by fold



Note. Steps on the x-axes correspond to the moments where the model has been trained on ACRIMA (step 1), ACRIMA and RIM-ONE-DL (step 2) and ACRIMA, RIM-ONE-DL and PAPILA (step 3). Notice also that the y-axis ranges from 0.5 to 1, instead of 0 to 1. This is to better visualize changes in score values.

Interestingly, the plots also show that sometimes re-training can hurt performance, even on the current dataset. At step 2 of the fourth fold, the model has been trained on ACRIMA and RIM-ONE-DL, and it already has a reasonable balanced accuracy on PAPILA ($BA = 0.637$ for naïve fine-tuning, $BA = 0.625$ for MAS), but if the network is fine-tuned, a reduction in performance is observed in PAPILA ($BA = 0.583$ for naïve fine-tuning, $BA = 0.56$ for MAS). The same happens in fold 2. If the model from step 2 is kept when training does not improve balanced accuracy in the last domain (PAPILA), then the results are the ones showed in Table 5. Compared with the values observed in Table 3 and 4, with this correction, absolute values tend to increase for both approaches, most noticeably in PAPILA scores. However, the differences between strategies remain roughly the same.

Table 5

Results for the compared continual learning strategies after completing all training steps but having stopped learning when necessary

Metric	**Strategy**	**Dataset**		
		ACRIMA	RIM-ONE-DL	PAPILA
Balanced accuracy	Naïve fine-tuning	0.827(0.024)	**0.829**(0.041)	**0.731**(0.064)
	MAS	**0.833**(0.019)	0.817(0.052)	0.71(0.058)
AUC	Naïve fine-tuning	0.9(0.009)	**0.912**(0.028)	**0.839**(0.067)
	MAS	**0.94**(0.002)	0.908(0.032)	0.835(0.059)

Note. Scores are averaged over the 4 folds and the standard deviation is written in parenthesis. For each dataset and metric, the highest value of the two compared approaches is highlighted in bold font. Statistically significant differences of balanced accuracy scores are indicated with one asterisk ($p < 0.05$), two ($p < 0.01$) or three ($p < 0.001$) next to the highest score.

4. Discussion

With the evidence at hand, it is possible to conclude that applying some form of sequential transfer learning, which is the core of the studied continual learning algorithms, seems to be the most sensible option to improve generalizability. Pre-processing techniques such as CLAHE can help, but the biggest gain in performance comes when the model can be trained with different datasets. Despite that, catastrophic forgetting is observed whenever the model is trained to learn several domains sequentially. The use of a continual learning approach designed specifically to deal with this effect, like MAS, does not seem to reduce the magnitude of forgetting. In fact, the naïve fine-tuning strategy apparently surpasses MAS in terms of performance. Yet, as statistical tests show, this advantage might not be considered statistically significant. This trend is observed for RIM-ONE-DL and PAPILA datasets, but not for ACRIMA, where MAS achieves a better result. Although the difference is not statistically significant, it is worth commenting that this finding could be related with the stability-plasticity dilemma. MAS is better suited to preserve old knowledge, but this comes at the expense of reducing plasticity, the ability to learn new patterns from newer data. Naïve fine-tuning, on the other hand, favors plasticity and thus tends to perform worse on older domains.

The non-significant difference between strategies are not completely surprising, even if correct hyper-parameter choice is assumed. In presence of high label noise, very similar images could have different labels even if the real diagnosis is the same. In an extreme case, the same fundus photo could be labelled differently by different experts. In this context, preserving the value of important parameters associated with some features of a domain, as MAS tries to do, is difficult because the same features, and therefore the same parameters, are important for the new domains.

For instance, in ACRIMA, there could be a relevant weight in the hidden layer of the classifier that is associated with a feature that is key to discriminate glaucoma from normal. Maybe the bigger the magnitude of that feature, the more likely the diagnosis of glaucoma. MAS would try to prevent changes in that parameter. However, if in RIM-ONE-DL greater values of that same feature tend to be present in normal cases, then an interference could happen because, despite the regularization, the value of the parameter associated with the mentioned feature would need to change to perform well on the RIM-ONE-DL.

In the Appendix B we show how sometimes images from different datasets and different diagnosis have more similar representations than images with the same labels. While not conclusive, this evidences that there might be a problem with labeling. This is also supported by the fact that the similarity phenomenon occurs between PAPILA images and images from the other two datasets. As was described in the method section, PAPILA researchers had access to clinical information of the patients, which means that their image annotation was not based only on information present in the retinographies.

4.1 Limitations

We recognize various limitations that constrain the extent of our work, so findings should be interpreted carefully.

First, hyper-parameter configuration may not have been chosen appropriately. Hyper-parameter tuning was done judiciously, but there are simply not enough observations to properly test all configurations. This is especially relevant for the case of MAS approach. The value of $\lambda = 1$ might not be optimal and other values could lead to better performance of the strategy.

Secondly, domain order was not considered. The learning sequence was fixed and the domains were learned in a pre-specified order. We arranged the datasets based on their publication dates, to resemble the most natural learning sequence, but there are five other possible ways to order the three datasets. It would have been interesting to study the influence of the ordering effect because it is known to have an impact in the context of continual learning, although not always very relevant (De Lange et al., 2022). With the computational resources available for the present project, however, testing all possible learning sequences involved a high cost difficult to meet.

It could also be argued that the CNN capacity was not enough to fully exploit the MAS method. While this is possibility is worth being explored thoroughly, it may not be the crucial limiting factor. We conducted separate experiments outside of this study to compare the two continual learning approaches, having more trainable layers, but we observed the same results.

Another limitation is that only two of the many possible continual learning approaches were compared. Because the goal of this line of work is to develop strategies that perform well in a domain incremental learning scenario, it is important for future research to evaluate more continual learning algorithms to determine their effectiveness in comparison with the ones tested here, with Learning Without Forgetting (Li & Hoiem, 2018) and Elastic Weight Consolidation (Kirkpatrick et al., 2017) being some options.

Last but not least, and closely related with the first point, the reduced sample size of datasets is an important cause of concern and reason to be cautious. The high variability found in cross-validation seems to indicate that random partition of the data does not generate completely equivalent sub-samples and this is in part due to the small sample size. With few observations, the effect on training of some outliers or hard-to-classify exemplars is greater.

4.2 Conclusions

The present work represents a contribution to the study of the generalization and continual learning capabilities of CNN models in an applied and highly relevant context such as glaucoma detection. The lack of generalization capability of models trained on datasets from a single source observed by other researchers have been confirmed. However, generalization can be improved applying pre-processing techniques such as CLAHE. It also has been shown that a continual learning perspective can be helpful to develop robust classifiers.

Two fine-tuning-based strategies have been compared, namely naïve fine-tuning and MAS, with results indicating that neither approach is superior to the other, despite MAS being specifically designed to prevent forgetting in a continual learning scenario. It should be noted that the study design had some limitations, and caution is advised in the interpretation of results. We hypothesize that the negative results obtained may be attributed in part to the high degree of label noise present in the datasets, which can make learning difficult. If our hypothesis is correct, then it would be advisable to try other forms of input that are more objective and less reliant on human judgement, such us OCT or VF tests.

5. References

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T. (2018). Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 139-154).

Aljundi, R., Kelchtermans, K., & Tuytelaars, T. (2019). Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11254-11263).

Bengio, Y., & Grandvalet, Y. (2003). No unbiased estimator of the variance of k-fold cross-validation. *Advances in Neural Information Processing Systems*, *16*.

Bouckaert, R. R., & Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In H. Dai, R. Srikant & C. Zhang (Eds), *Advances in Knowledge Discovery and Data Mining* (pp. 3-12). Springer.

Batista, F. J. F., Diaz-Aleman, T., Sigut, J., Alayon, S., Arnay, R., & Angel-Pereira, D. (2020). Rim-one dl: A unified retinal image database for assessing glaucoma using deep learning. *Image Analysis & Stereology*, *39*(3), 161-167. https://doi.org/10.5566/IAS.2346

Casson, R. J., Chidlow, G., Wood, J. P., Crowston, J. G., & Goldberg, I. (2012). Definition of glaucoma: clinical and experimental concepts. *Clinical & experimental ophthalmology*, *40*(4), 341-349. https://doi.org/10.1111/j.1442-9071.2012.02773.x

Chen, D., Anran, E., Tan, T. F., Ramachandran, R., Li, F., Cheung, C., Yousefi, S., Tham, C. C. Y., Ting, D. S. W., Zhang, X., & Al-Aswad, L. A. (2023). Applications of Artificial Intelligence and Deep Learning in glaucoma. *Asia-Pacific Journal of Ophthalmology*, *12*(1), 80-93. https://doi.org/10.1097/APO.0000000000000596

De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2022). A continual learning survey: defying forgetting in Classification Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(7), 3366–3385. https://doi.org/10.1109/TPAMI.2021.3057446

Derakhshani, M. M., Najdenkoska, I., van Sonsbeek, T., Zhen, X., Mahapatra, D., Worring, M., & Snoek, C. G. (2022, September). LifeLonger: A Benchmark for Continual Disease Classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II* (pp. 314-324). Cham: Springer Nature Switzerland.

Dey, A., & Bandyopadhyay, S. K. (2016). Automated glaucoma detection using support vector machine classification method. *British Journal of Medicine and Medical Research*, *11*(12), 1-12.

Díaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J. M., & Navea, A. (2019). CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online*, *18*, 1-19. https://doi.org/10.1186/s12938-019-0649-y

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, *3*(4), 128-135.

Gómez-Valverde, J. J., Antón, A., Fatti, G., Liefers, B., Herranz, A., Santos, A., Sánchez, C. & Ledesma-Carbayo, M. J. (2019). Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. *Biomedical optics express*, *10*(2), 892-913.

Gonzalez, R. & Woods, R. (2018). Intensity transformations and spatial filtering. In R. Gonzalez & R. Woods, *Digital image processing* (4ed), (pp. 119-203). Pearson education.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Introduction. In I. Goodfellow, Y. Bengio & A. Courville, *Deep Learning* (pp.1-29). MIT Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Regularization for Deep Learning. In I. Goodfellow, Y. Bengio & A. Courville, *Deep Learning* (pp.221-267). MIT Press.

Gupta, N., Garg, H., & Agarwal, R. (2021). A robust framework for glaucoma detection using CLAHE and EfficientNet. *The Visual Computer*, 1-14.

Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. (2020). Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, *24*(12), 1028-1040. https://doi.org/10.1016/j.tics.2020.09.004

Han, J., Kamber, M., & Pei, J. (2012). Data preprocessing. In J. Han, M. Kamber & J. Pei, *Data mining concepts and techniques* (3ed), (pp. 83-125). Morgan Kauffman.

Hsu, Y. C., Liu, Y. C., Ramasamy, A., & Kira, Z. (2018). Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*.

Iman, M., Arabnia, H. R., & Rasheed, K. (2023). A review of deep transfer learning and recent advancements. *Technologies*, *11*(2), 40. https://doi.org/10.3390/technologies11020040

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *International Conference on Machine Learning* (Vol. 1, pp. 448–456). http://ece.duke.edu/~lcarin/Zhao12.17.2015.pdf

Jung, H., Ju, J., Jung, M., & Kim, J. (2018). Less-Forgetful Learning for Domain Expansion in Deep Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence, 32*(1). https://doi.org/10.1609/aaai.v32i1.11769

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, *114*(13), 3521-3526. https://doi.org/10.1073/pnas.1611835114

Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., & Sancho-Gómez, J. L. (2022). PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, *9*(1), 291.https://doi.org/10.1038/s41597-022-01388-1

Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence, 40*(12), 2935-2947. https://doi.org/10.1109/TPAMI.2017.2773081

Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A. D., & Van De Weijer, J. (2022). Class-Incremental Learning: Survey and Performance Evaluation on Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 1–20. https://doi.org/10.1109/tpami.2022.3213473

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation,* 4, 109-165. Academic Press.

Mirzania, D., Thompson, A. C., & Muir, K. W. (2021). Applications of deep learning in detection of glaucoma: a systematic review. *European Journal of Ophthalmology, 31*(4), 1618-1642. https://doi.org/10.1177/1120672120977346

Murphy, K. (2022). Learning with fewer labeled examples. In K. Murphy, *Probabilistic machine learning: an introduction* (pp. 625-626). MIT Press.

Nadeau, C., & Bengio, Y. (1999). Inference for the generalization error. *Advances in neural information processing systems, 12.*

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems, 32, 8026-8037.* https://arxiv.org/pdf/1912.01703.pdf

Pytorch. (2023). *Image classification reference training scripts*. Retrieved April 26, 2023, from https://github.com/pytorch/vision/tree/main/references/classification

Raschka, S. (2018). Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808.*

van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Ruder, S. (2017). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747.*

Russ, J. C. (2011). Image enhancement in the spatial domain. In J.C. Russ, *The image processing handbook* (6ed) (pp.269-337). CRC Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C. & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision, 115*, 211-252.

Schwarz, J., Czarnecki, W., Luketina, J., Grabska-Barwinska, A., Teh, Y. W., Pascanu, R., & Hadsell, R. (2018, July). Progress & compress: A scalable framework for continual learning. In *International conference on machine learning* (pp. 4528-4537). PMLR.

Shen, D., Wu, G., & Suk, H. I. (2017). Deep learning in medical image analysis. *Annual review of biomedical engineering*, *19*, 221-248.https://doi.org/10.1146/annurev-bioeng-071516

Shibata, N., Tanito, M., Mitsuhashi, K., Fujino, Y., Matsuura, M., Murata, H., & Asaoka, R. (2018). Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Scientific reports*, *8*(1), 14665. https://doi.org/10.1038/s41598-018-33013-w

Shoukat, A., Akbar, S., Hassan, S. A. E., Rehman, A., & Ayesha, N. (2021, December 13-14). *An automated deep learning approach to diagnose glaucoma using retinal fundus images*. International Conference on Frontiers of Information Technology (FIT) 2021, Islamabad, Pakistan. https://doi.org/10.1109/FIT53504.2021.00031

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Van De Ven, G. M., Tuytelaars, T., & Tolias, A. S. (2022). Three types of incremental learning. *Nature Machine Intelligence*, *4*(12), 1185–1197. https://doi.org/10.1038/s42256-022-00568-3

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, *27*, 3320-3328.

Zhang, N., Wang, J., Li, Y., & Jiang, B. (2021). Prevalence of primary open angle glaucoma in the last 20 years: a meta-analysis and systematic review. *Scientific Reports*, *11*(1), 13762. https://doi.org/10.1038/s41598-021-92971-w

Zuiderveld, K. (1994). Contrast limited adaptive histogram equalization. *Graphics gems*, 474-485. https://doi.org/10.1016/b978-0-12-336156-1.50061-6

6. Appendix A

MAS implementation with Python

```
import torch

# tfm is a custom package with some utility functions and classes
from tfm import ImgData
import torch.nn as nn

def compute_importance(model:nn.Module,
                      data:ImgData,
                      device:str,
                      normalize:bool = True) -> dict:
  """
  Importance estimation as described in the paper
  https://openaccess.thecvf.com/content_ECCV_2018/papers/Rahaf_Aljundi_Memory_
Aware_Synapses_ECCV_2018_paper.pdf

  """

  grads = {name:torch.zeros(param.shape,device=device) \
        for name,param in model.named_parameters() if param.requires_grad}
  N = len(data)

  for X,y in data:
    X = X.unsqueeze(0).to(device)
    model.eval()
    output = torch.sigmoid(model(X)) # compute output of learned function
    model.zero_grad() # reset gradient to zero
    output.backward() # compute gradient of output
    # Gradients are accumulated over the given data points to obtain
importance weight Ωij for parameter θij
    for name,param in model.named_parameters():
      if param.requires_grad:
        grads[name] += param.grad.abs() # update gradient for each param

  # average gradient is obtained for each parameter
  omega = {name:sum_grad/N for name,sum_grad in grads.items()} # calculate
gradient mean

  # vector normalization with l2 norm
  if normalize:
    omega_norm = torch.sqrt(sum([mean_grad.pow(2).sum() for name,mean_grad in
omega.items()]))
    omega = {name:mean_grad/omega_norm for name,mean_grad in omega.items()}
  return omega


def update_importance(prev_omega,
                      model:nn.Module,
                      data:ImgData,
                      device:str,
                      normalize:bool = True) -> dict:
  """
  Importance updating as described in the paper
  https://openaccess.thecvf.com/content_CVPR_2019/papers/Aljundi_Task-
Free_Continual_Learning_CVPR_2019_paper.pdf

  """
  # compute Ω for new domain
  curr_omega = compute_importance(model,data,device,normalize=True)
  # cumulative moving average
  for name in curr_omega.keys():
    curr_omega[name] = (prev_omega[name] + curr_omega[name])/2

    return curr_omega



class MASLoss(nn.Module):
  """
  Complete loss function as described in the paper
  https://openaccess.thecvf.com/content_ECCV_2018/papers/Rahaf_Aljundi_Memory_
Aware_Synapses_ECCV_2018_paper.pdf

  """
  def __init__(self,
              model:nn.Module,
              lambd:float,
              importance:dict,
              loss_function:nn.Module) -> None:

    super().__init__()
    self.model = model
    self.old_params = {name:param.data.detach().clone() for name,param in
model.named_parameters() if param.requires_grad}
    self.lambd = lambd # regularization hyper-parameter λ
    self.omega = importance
    self.loss = loss_function

  def forward(self,
             input:torch.Tensor,
             target:torch.Tensor) -> torch.Tensor:
```

```
reg = 0. # regularization term
for name,param in self.model.named_parameters():
  if param.requires_grad:
    reg += (self.omega[name]*(param - self.old_params[name])**2).sum()

return self.loss(input,target) + self.lambd*reg
```

7. Appendix B

Complementary analysis

An additional analysis was performed to explore the causes of the observed outcomes in the main experiment comparing continual learning strategies. The fixed feature extractor of the network was used to get the representations of all the images in the three datasets. Then, an average representation or prototype of each domain and class was computed by taking the mean over all the exemplars grouped by dataset and label. Similarity between all possible pairs of prototype vectors was calculated using cosine similarity. The resulting similarity matrix can be seen in Table 6.

The table shows that the similarity between the average representation of a ACRIMA glaucomatous image and a PAPILA normal image (0.958) is higher than the similarity between two images labeled as glaucoma in both datasets (0.946). The same phenomenon occurs with RIM-ONE-DL and PAPILA images, although the difference is smaller (difference of 0.02).

Table 6

Similarity matrix of average representations

		Normal			**Glaucoma**		
		A	R	P	A	R	P
Normal	A	1	-	-	-	-	-
	R	0.949	1	-	-	-	-
	P	0.969	0.97	1	-	-	-
Glaucoma	A	0.973	0.939	0.958	1	-	-
	R	0.933	0.957	0.953	0.941	1	-
	P	0.940	0.935	0.959	0.946	0.951	1

Note. For space reasons, dataset names are shortened. A stand for ACRIMA, R for RIM-ONE-DL, and P for PAPILA. Values in cells correspond to pairwise cosine similarities, which range from -1 to 1.