Evaluating the performance of hyperparameters for unbiased and fair machine learning

Vy Bui[a], Hang Yu[a], Karthik Kantipudi[b], Ziv Yaniv[b], and Stefan Jaeger[a]

[a]National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA
[b]National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892, USA

ABSTRACT

Biased outcomes in machine learning models can arise due to various factors, including limited training data, imbalanced class distribution, suboptimal training methodologies, and overfitting. In training neural networks with Stochastic Gradient Descent (SGD) and backpropagation, the choice of hyperparameters like learning rate and momentum is crucial to influencing the model's performance. A comprehensive grid search study was conducted using static hyperparameters with standard SGD and dynamic hyperparameters with the Adam optimizer. The investigation focused on a multifaceted analysis across different tasks —- classification, segmentation, and detection — and was applied to four image-based applications: digit classification using the MNIST dataset, tuberculosis detection from chest X-ray images, lung segmentation in chest X-ray images, and the detection of malaria parasites in blood smear images. In the first comparative study on the MNIST dataset, the SGD algorithm consistently outperformed the Adam algorithm as noise levels increased. SGD held a slight advantage in accuracy over Adam in a noise-free environment. This advantage became more apparent as noise was introduced and increased to moderate levels. At high noise levels, both algorithms experienced a significant decline in performance, yet SGD maintained a relatively better accuracy compared to Adam. This trend underscored SGD's superior ability to generalize across varying noise conditions. For TB detection in the second application, the DenseNet121 architecture was used, and it was found that Adam showed better performance on the larger TBX11K dataset. However, SGD outperformed Adam on a smaller subset of TBX11K. In the third comparative study for lung segmentation using COVID-19 and NLM datasets, SGD slightly outperformed Adam based on the mean and Hausdorff distances. In the case of malaria parasite detection using the YOLOv8 architecture, SGD and Adam optimizers showed varying performances across different conditions. Initially, both achieved high accuracies on the full dataset, with Adam slightly outperforming SGD. However, with decreasing dataset size, SGD maintained more consistent performance, while Adam's accuracy fluctuated significantly. In noise tests, both showed equal accuracy on clean data, but under noise conditions, SGD maintained higher accuracies, suggesting better generalization capabilities. Further experiments assessed how well SGD and Adam could cope with domain shifts, mainly when using data from different countries. SGD's generalization performance was superior to Adam's performance in these experiments. In summary, although Adam is arguably the more popular of the two optimization techniques, SGD held its own in the experiments and showed superior performance when it came to generalization or classifying in noisy conditions.

Keywords: Unbiased machine learning, fair machine learning, optimization, grid search, stochastic gradient descent, hyperparameters, SGD, Adam, MNIST digits classification, TB detection, lung segmentation, chest X-ray, malaria parasite detection

1. INTRODUCTION

Biased machine learning models can arise from insufficient training data or a biased distribution of class samples in the training data, but also because of inadequate training methods, and over-training.[1] Generalization refers to the ability of a machine learning model to perform well on new, unseen data that was not included in the training set. Memorization in this context refers to the model's ability to learn and remember the training data.

Send correspondence to stefan.jaeger@nih.gov or vy.bui@nih.gov.

Medical Imaging 2024: Image Processing, edited by Olivier Colliot, Jhimli Mitra, Proc. of SPIE Vol. 12926, 1292614 · © 2024 SPIE
1605-7422 · doi: 10.1117/12.3006909

While good memorization is necessary to achieve low training error, excessive memorization or overfitting can impede generalization.

For training neural networks with stochastic gradient descent and backpropagation, several hyperparameters affect the performance of the trained model. By conducting a comprehensive grid search, the generalization performance of networks trained with different static and dynamic hyperparameters, specifically learning rate (LR) and momentum (M), is investigated. For this purpose, static hyperparameters with the standard stochastic gradient descent (SGD) optimizer and dynamic hyperparameters with Adaptive Moment Estimation (Adam) are employed. SGD and Adam are two of the most popular optimization strategies for neural networks. Since its first appearance,[2] Adam has often been chosen over SGD due to its ability to converge more quickly toward the minimum of a loss function, leading to a more efficient training process with lower training losses. This advantage is primarily attributed to Adam's adaptive learning rate mechanism, which allows it to make informed updates to the model parameters by considering both the first and second moments of the gradients. However, despite Adam's superior performance during the training phase, some authors observed that models optimized with SGD tend to exhibit better generalization when evaluated on unseen test data. For example, SGD displayed better generalization ability than adaptive optimization methods like Adam in.[3–8] Despite these observations, Adam remains widely used as the main optimization tool in deep learning.[9–13]

Quantifying the effect of different hyperparameters on a network's performance helps to measure generalizability. Understanding hyperparameter sensitivity is essential for optimizing machine learning models, and tuning parameters such as learning rate and momentum is crucial for efficiently training machine learning models. In this work, an extensive grid search study compared static hyperparameters for the standard SGD and dynamic hyperparameters for the Adam optimizer based on the following four tasks: handwritten digit classification (MNIST dataset), tuberculosis positive or negative classification (chest X-ray images), lung segmentation (chest X-ray images), and malaria parasite detection (blood smear images). The experiments presented in this study are a step toward a better understanding of the requirements for unbiased and fair machine learning. The study emphasizes the critical role of selecting the right hyperparameters to enhance model generalization, thereby contributing to the ongoing effort to refine and improve the generalization ability of machine learning models in practical applications.

2. METHODS

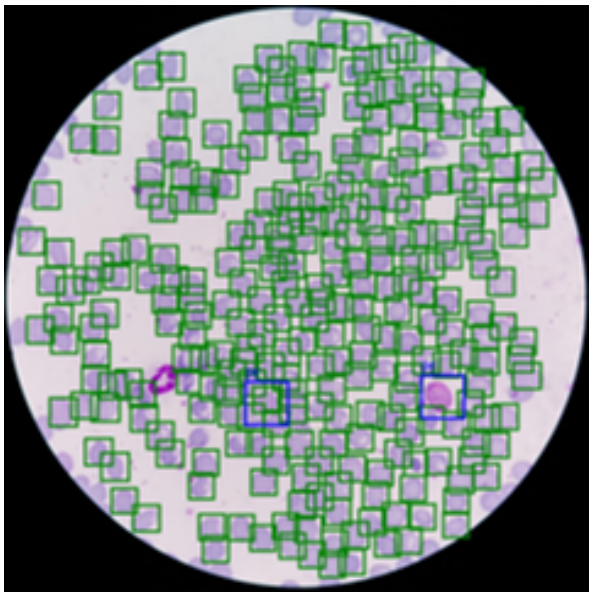
This study conducted a thorough assessment of the generalization capabilities of several neural network architectures, examining the effects of both static and dynamic hyperparameters. Specifically, the study focused on the two widely used optimizers, SGD and Adam, and two key hyperparameters, the learning rate and momentum. The aim was to gain a comprehensive understanding of how the optimizer and its hyperparameters impact the overall ability of a network to generalize to unseen data. The experiments included four distinct tasks: 1) classification of handwritten numerical digit images using a custom convolutional neural network, 2) classification of chest X-ray images for tuberculosis using the DenseNet121 network, 3) semantic segmentation of the lungs in chest X-ray images using the YOLOv8m-seg network, and 4) detection of cells in microscopy images of thin blood smears using the YOLOv8x network. Figure 1 shows sample images from the datasets used in this work.

The grid search experiments conducted in this study were extensive, encompassing a wide range of LR and M values to explore their impact on model performance. Specifically, the experiments used six distinct LR values: 0.0001, 0.001, 0.01, 0.016, 0.1, and 0.2. These values were chosen to cover a broad spectrum, from very low to relatively high, to understand how the models respond to each LR. Note that the specific learning rate of 0.016 was included as proposed in.[1] Alongside these learning rates, ten momentum values were used: 0.0, 0.2, 0.4, 0.6, 0.8, 0.825, 0.85, 0.874, 0.9, and 0.925, where the momentum value of 0.874 was also proposed in.[1] By exploring this comprehensive set of learning rates and momentum values, the experiments aimed to provide a better understanding of how these hyperparameters affect the training process and overall performance of the models. In the case of the Adam optimizer, the momentum was beta1, while beta2 was set to 0.999 for all experiments.

(a) MNIST

(b) TBX11K

(c) COVID19



(d) NLM Malaria

Figure 1: Sample images from the datasets used in this work. (a) handwritten digit images, (b) frontal chest X-ray images with associated TB/not-TB labels, (c) frontal chest X-ray images with associated manual ground-truth lung segmentations, and (d) blood smear images with bounding boxes denoting infected and uninfected cells.

2.1 Handwritten Digit Classification

For digit classification, the optimization performances of SGD and Adam were evaluated under varying noise levels, i.e., 5%, 10%, 15%, and 20%, using the MNIST dataset. These percentages represent the proportion of randomly inverted pixels (intensity modified to 255 minus original intensity). This alteration was only applied to the test set, leaving the training and validation sets untouched.

A deep learning model based on an 11-layer convolutional neural network (CNN) was trained. The CNN model started with a convolutional layer that took a single-channel input (grayscale image) and applied 16 filters, followed by a second convolutional layer that expanded the channel size to 32. Both convolutional layers used a kernel size of 3 with stride 1 and padding 1. After each convolution, a ReLU activation function introduced non-linearity, and a max pooling operation with a 2×2 kernel and stride reduced the spatial dimensions by half. A dropout layer with a rate of 0.25 was applied after flattening the output to prevent overfitting. The network concluded with two fully connected layers with a final output of 10 classes. The maximal output value determined the final class of a given image. The number of parameters was estimated at around 200 thousand for an input image size of 28×28. Weight initialization was performed using the Kaiming uniform method.[14] No data augmentation techniques were applied; however, the input was normalized to the range $[-1; 1]$. The training used a batch size of 64 and was conducted over 30 epochs, employing cross entropy as the loss function. Sizes of the training, validation, and test datasets were 54,000, and 6,000, and 10,000, respectively. The model's performance was assessed through 10-fold cross-validation. Fig. 2 shows examples of both clean data and data with varying noise levels.

Figure 2: Visual Comparison of the MNIST Dataset at different noise levels. The figure displays sample images from the MNIST dataset, showcasing an original clean image alongside sample images with progressively increasing noise levels of 5%, 10%, 15%, and 20% achieved through random pixel inversion.

2.2 Tuberculosis Classification

In this task, TB and non-TB cases were classified using chest X-ray images from the TBX11K dataset (1,600 images).[15,16] To assess the performances of SGD and Adam with different learning rates and momentums, a series of experiments using both optimizers were conducted. The experiments evaluated the effectiveness of SGD and Adam for tuberculosis classification, with a focus on varying training set sizes. First, a baseline was established by comparing the performance of each optimizer using the complete dataset. To further explore their performance, particularly in situations with limited data availability, the dataset was incrementally reduced. The experiments were repeated with training samples comprising 75%, 50%, and 25% of the original dataset size, respectively. This stepwise reduction in dataset volume provided insights into how each optimizer behaves for different dataset sizes. DenseNet121 was employed for all experiments, and no pre-trained weights were used; weight initialization was performed using the Kaiming uniform method.[14] Images were resampled to 256×256. No data augmentation was applied. The input was normalized to the range [-1;1]. Training used a batch size of 32 and was conducted over 100 epochs, employing cross entropy as the loss function. The sizes of the training, validation, and test datasets were 1,024, 257, and 321 images, respectively. The model's performance was assessed through 5-fold cross-validation, with standard accuracy used to measure the performance of detecting TB and non-TB cases.

2.3 Chest X-ray Lung Segmentation

The performance of SGD and Adam was analyzed for a lung segmentation task on chest X-ray images. The first part of this task involved training and testing using a COVID-19 dataset,[17] where the training set consisted of 2,963 images, with a validation set of 977 images to fine-tune the model parameters, and a separate test set of 1,301 images. Both SGD and Adam optimizers underwent the same training and validation process for

a direct performance comparison on the COVID-19 dataset. The study was further extended to evaluate the performance on the NLM dataset,[18] which combines the Montgomery and Shenzhen datasets,[19,20] containing normal and abnormal chest X-rays with manifestations of TB. A model was trained on 337 images, validated on 112 images, and finally tested on 141 images, using the YOLOv8m-seg architecture.[21] COCO weights were used for model initialization. Images were resampled to 640×640. Training used a batch size of four and was conducted over 100 epochs. The performance of SGD and Adam was evaluated based on the Dice coefficient, mean distance, and Hausdorff distance metrics.

2.4 Malaria Cell Detection

In the last task, a YOLOv8x[21] was trained to detect Plasmodium falciparum (P. falciparum) and Plasmodium vivax (P. vivax), the most common malaria parasite species, in thin blood smear images.[22,23] In particular, the models were trained and evaluated using a dataset of 364 patient cases acquired in Bangladesh, comprising 3,532 images, 7,952 instances of P. falciparum, 4,346 instances of P. vivax, and over 860,000 instances of uninfected cells. Another dataset with 190 patients was acquired in Sudan,[24] comprising 874 images, 336 instances of P. falciparum, 59 instances of P. vivax, and over 220000 instances of uninfected cells. The YOLOv8x model was initialized using COCO weights. Images were resampled to 1024×1024. Training used a batch size of 30 and was conducted over 500 epochs.

The experiments were first designed to assess the performance of SGD and Adam across varying dataset sizes in the context of malaria cell detection. This set a benchmark for understanding how each optimizer performs under full-data conditions. To delve deeper into the optimizers' capabilities, especially in scenarios with limited data, the dataset size was systematically reduced. The experiments were replicated with training samples constituting 75%, 50%, and 25% of the full-size dataset. This degradation in dataset size allowed observing how each optimizer performs under varying amounts of data.

Furthermore, the study was expanded to include experiments designed to evaluate the optimizers' generalization performance for domain-shift challenges. The first domain-shift experiment was done by introducing various types of noise to the Bangladesh test set. Specifically, the models were tested on the Bangladesh data altered with Gaussian, Poisson, and salt-and-pepper noise. These types of noise were chosen to represent common image degradation scenarios. The inclusion of noise during testing allowed assessing the generalization performance and noise resilience of SGD and Adam. The second domain-shift experiment involved training models on data from Bangladesh and subsequently testing them on data from Sudan. This setup aimed to evaluate the optimizers' robustness and adaptability across different data sources.

The evaluation metrics used depended on the availability of ground truth data. When the ground truth comprised bounding boxes, the mean average precision at 50 (mAP50) was used. mAP50 is a performance measure that considers both precision and recall across different levels of confidence thresholds. This metric is particularly useful for assessing object detection tasks. In situations where the ground truth was limited to cell counts, as seen in the Sudan data, the evaluation was based on accuracy. The accuracy metric quantifies the disparity between predicted and actual cell counts, providing a clear indication of the model's performance in estimating quantities.

3. RESULTS

3.1 Handwritten Digit Classification

The comparative analysis of SGD and Adam at various noise levels, as depicted in Fig. 3, showed the relative performance of each optimizer under noisy test conditions for the MNIST dataset. The reported accuracies represented the mean performance of the top 10 models identified in the grid search experiment. At the baseline condition of 0% noise, SGD exhibited marginally better performance, achieving an accuracy of 99.33% as opposed to Adam's 99.28%. As noise is introduced (5% noise), the gap widened with SGD maintaining a higher accuracy of 97.46%, while Adam's performance dropped to 93.18%. The disparity became more pronounced for 10% noise, where SGD maintained a larger lead with an accuracy of 86.98%, showing a substantial margin over Adam, which fell to 73.18%. This trend of SGD outperforming Adam continued for 15% noise, with SGD's accuracy at 66.08% in contrast to Adam's 49.52%. Finally, for 20% noise, both optimizers showed a significant drop in performance;

SGD achieved an accuracy of 45.93%, while Adam's performance further declined to 32.53%. These findings showed that SGD's generalization performance is superior compared to Adam's. When facing increased noise levels, SGD consistently outperformed Adam.

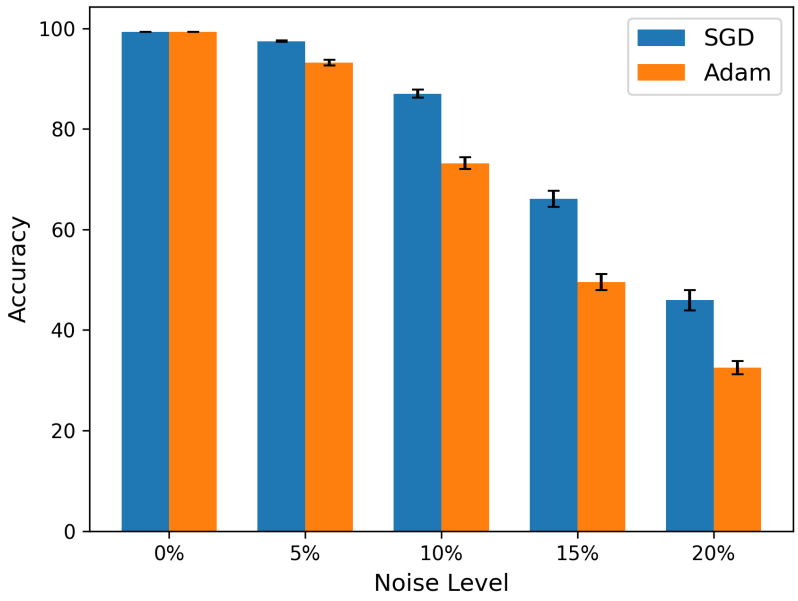


Figure 3: The chart compares the accuracy of models using SGD (blue) and Adam (orange) optimizers for the MNIST dataset across different noise levels (0% (clean), 5%, 10%, 15%, and 20%). Error bars indicate the variability among the top 10 performing models in the grid search experiment.

3.2 Tuberculosis Classification

Fig. 4 shows the best performances for SGD and Adam when classifying TB and non-TB cases on chest X-ray images from the TBX11K dataset. The reported accuracies were the average from the top 10 performing models in the grid search experiment. The accuracies differed only slightly. At full training set size (100%), SGD marginally outperformed Adam with 99.52% accuracy compared to Adam's 99.43%. At 75% training size, Adam edged SGD out with 99.27% accuracy against SGD's 99.24%. At 50%, SGD showed slightly higher accuracy than Adam, with SGD achieving 99.18% compared to Adam's 98.95%. At 25%, Adam showed slightly higher accuracy than SGD, which achieved 96.32% compared to Adam's 96.74%. These findings indicated that both algorithms demonstrate comparably high accuracy across different training set sizes, with no significant differences in their performances.

3.3 Chest X-ray Lung Segmentation

For lung segmentation using a COVID-19 dataset, the two optimizers showed a similar performance. Among the top 10 performing models, both SGD and Adam achieved an average Dice coefficient of 0.945. However, there were slight differences in the mean and Hausdorff distances, with SGD showing marginally better performance. The mean distance for SGD was 10.71 pixels compared to 10.80 pixels for Adam. The Hausdorff distance was 60.77 pixels for SGD, which was slightly lower than Adam's 61.72 pixels.

In another experiment, using the NLM dataset for lung segmentation, SGD and Adam provided closely matching results as well. SGD achieved an average Dice coefficient of 0.950, a mean distance of 21.93 pixels, and a Hausdorff distance of 142.71 pixels. In comparison, Adam demonstrated a nearly equivalent level of accuracy

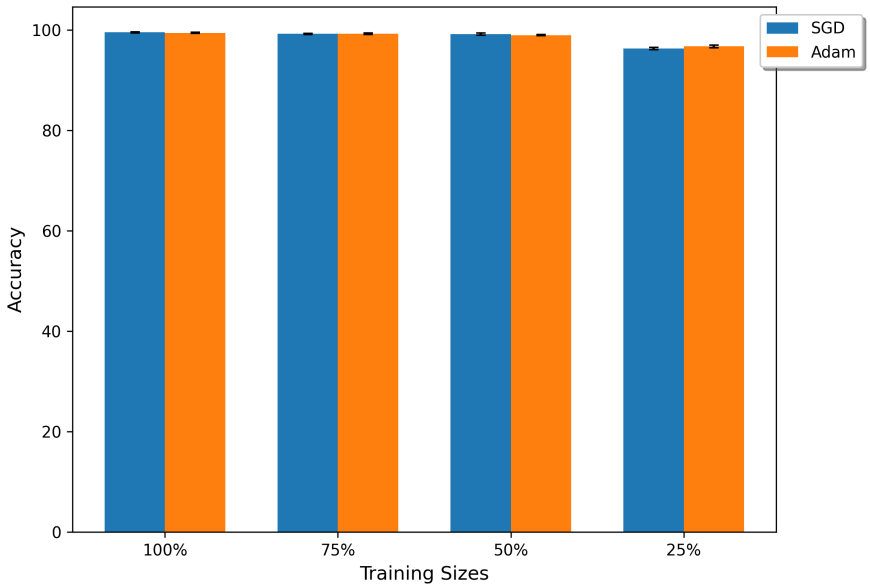


Figure 4: Accuracy of TB classification using SGD (blue) and Adam (orange) optimizers on the TBX11K dataset across different dataset sizes (100%, 75%, 50%, and 25%). Error bars indicate the variability among the top 10 performing models in the grid search experiment.

with an average Dice coefficient of 0.949. However, it showed a slightly less favorable performance in spatial distance measures, with a mean distance of 22.47 pixels and a Hausdorff distance of 146.72 pixels.

In summary, both SGD and Adam were highly effective for lung segmentation on the COVID19 and NLM dataset; however, SGD was marginally superior in terms of the mean and Hausdorff distances.

3.4 Malaria Cell Detection

The last experiments involved detecting and identifying P. falciparum, P. vivax, and uninfected blood cells in thin blood smear images. All the reported mAP50 values reflected the average performance of the top 10 models identified in the grid search experiment. First, SGD and Adam were evaluated on the malaria set across different dataset sizes, as shown in Fig. 5, where distinct performance patterns were observed for each optimizer. SGD and Adam started with high mean accuracies on the full-size dataset, with Adam (0.939) slightly outperforming SGD (0.938). However, as the dataset size decreased, the behaviors of the two optimizers changed. When the dataset was reduced to 75%, SGD maintained a relatively stable performance, with a mean accuracy of 0.900, while Adam's performance dropped significantly to 0.843. A similar behavior was observed for a dataset size of 50%, where SGD's mean accuracy slightly decreased to 0.888, and Adam's accuracy was 0.858. Surprisingly, for the smallest dataset size (25%), the mean accuracy of SGD increased to 0.898, and the accuracy of Adam increased to 0.899. These side-by-side comparisons showed that SGD and Adam performed almost identically on the full data; however, SGD showed more consistent performance across different dataset sizes, demonstrating greater stability with significant data reductions.

In a second experiment, SGD and Adam were evaluated under different noise conditions in the malaria test set. Again, different behaviors were observed, as shown in Fig. 6. SGD and Adam achieved the same mean accuracy of 0.939 for the clean (no noise) test set, demonstrating again equal performance under ideal conditions. However, their performance was different when Gaussian noise was introduced: SGD maintained a relatively high accuracy of 0.900, while Adam's accuracy dropped to 0.834. This pattern persisted when Poisson noise was used instead, where SGD achieved a mean accuracy of 0.885, outperforming Adam's 0.810. The disparity in their

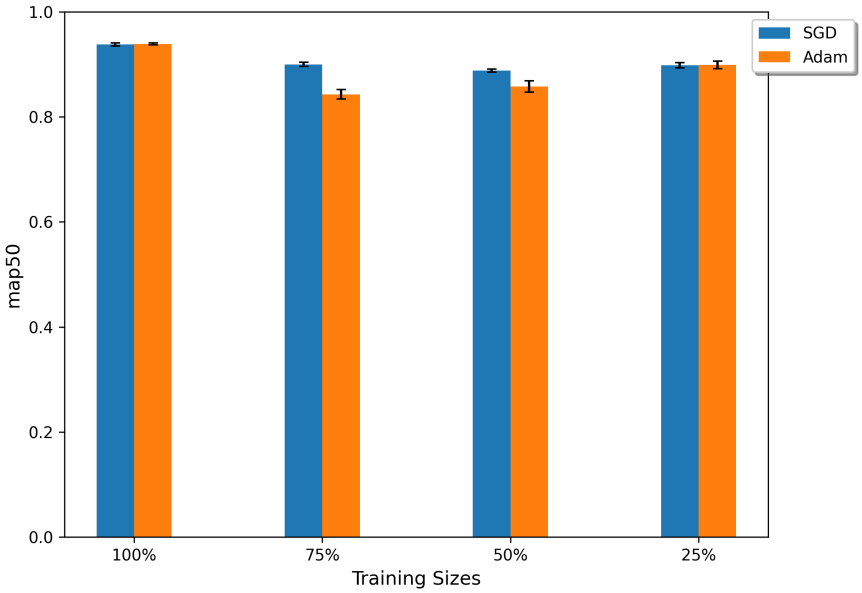


Figure 5: mAP50 accuracy of the YOLOv8x detection model using SGD (blue) and Adam (orange) for the malaria dataset across different dataset sizes (100%, 75%, 50%, and 25%). Error bars indicate the variability among the top 10 performing models in the grid search experiment.

performance became even more pronounced with salt-and-pepper noise, with SGD achieving a mean accuracy of 0.823 compared to Adam's 0.730. These findings suggested again that SGD had better generalization capabilities across varied noise conditions, maintaining consistent performance. In contrast, Adam appeared to be more prone to memorizing the training data, leading to a steeper decline in accuracy when exposed to noise variations in the test set.

The final experiment assessed the ability of SGD and Adam to handle domain-shift challenges arising from different data sources (countries). This was achieved by training the models on data from Bangladesh and then evaluating their performance on data from Sudan. SGD's best performance in detecting infected cells was 75.96%, while Adam achieved a best performance of 71.03%. Thus, SGD again demonstrated its superior generalization performance over Adam for the domain-shift problem.

3.5 Optimal Learning Rate and Momentum Configurations for SGD and Adam

For the experiments above, Table 1 summarizes the best-performing learning rates and momentum values across all datasets for different variations of dataset size and noise levels. The reported LR and M values produced the best-performing models in the grid search experiments.

3.5.1 SGD

For the MNIST dataset, optimal performance on clean data was achieved with a learning rate and momentum combination of (0.016, 0.825). When the data contained 5%, 10%, and 15% noise, the best results were observed for the pair (0.01, 0.4). At the higher noise level of 20%, the most effective setting was (0.0001, 0.4). Thus, the optimal results were obtained with a learning rate of around 0.01/0.016 for most noise levels. Only for heavy noise (20%), a lower learning rate performed better.The best momentum for SGD was 0.825 for 0% noise and then changed to 0.4 for all other noise levels.

For tuberculosis classification using the TBX11K dataset, the most effective learning rate and momentum combination for 100% training set size was (0.01, 0.925). For 75%, the optimal pairing was found to be (0.016,

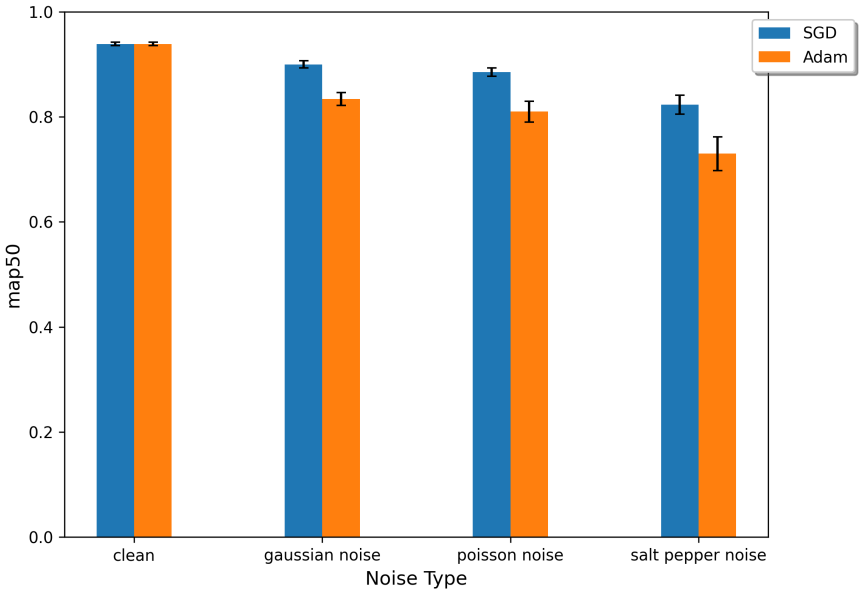


Figure 6: Comparative analysis of the mAP50 accuracy achieved by the YOLOv8x detection model using two different optimizers, SGD (blue) and Adam (orange), when applied to the malaria dataset. The comparison spans various noise conditions in the test set, including a clean (no noise) scenario, Gaussian noise, Poisson noise, and salt-and-pepper noise. Error bars are included to illustrate the variability observed among the top 10 performing models in the grid search experiment under each noise condition.

0.8). Furthermore, for a training size of 50%, the best-performing pair was (0.016, 0.9), and for the smallest training set size of 25%, the ideal configuration was (0.1, 0.4). These results suggested that moderate learning rates around 0.01/0.016 coupled with strong momentum 0.8/0.9/0.925 were performing better for larger datasets (100%-50%). On the other hand, in the case of a smaller dataset (25%), there was a preference for a higher learning rate of 0.1 combined with a smaller momentum of 0.4.

For lung segmentation using COVID-19 datasets, different optimal pairings of learning rate and momentum were identified for the three evaluation metrics. For the Dice coefficient, the best performance was achieved with a pairing of (0.01, 0.4). In terms of mean distance, the most effective combination was (0.1, 0.0). For the Hausdorff distance, the optimal pairing was again (0.01, 0.4). These combinations indicated that SGD performed well with lower momentum values (0.0 or 0.4) across all metrics, with learning rates of 0.01 or 0.1, depending on the metric.

When using NLM datasets for the lung segmentation task, the Dice coefficient saw its best results with a pairing of (0.016, 0.6). For the mean distance metric, the most effective combination identified was (0.016, 0.2). In the case of the Hausdorff distance, the pairing that yielded optimal results was (0.001, 0.6). Thus, the COVID-19 and NLM datasets showed a tendency toward moderate and lower momentum values for different learning rates.

For the malaria task, with 100% training set size, the best (learning rate, momentum value) pairing was (0.01, 0.8). Furthermore, for a training set size of 75% it was (0.016, 0.825), for 50% it was (0.1, 0.6), and for the smallest size of 25%, it was (0.016, 0.4). These results showed that learning rates around 0.01/0.016 provided again the best performance, with a tendency toward smaller momentum values for smaller training sets.

Under different noise conditions, the following pairings were performing the best: For noise-free data, (0.01, 0.8) was optimal. With Gaussian noise, the pair (0.01, 0.85) was found optimal. For Poisson and salt-and-pepper

OPTIMIZER	TASK (DATASET)	LEARNING RATE, MOMENTUM
SGD	Digit Classification (MNIST noise levels)	(0.016, 0.825) (0.01, 0.4) (0.01, 0.4) (0.0001, 0.4)
	TB Classification (TBX11K dataset sizes)	(0.01, 0.925) (0.016, 0.8) (0.016, 0.9) (0.1, 0.4)
	Lung Segmentation (COVID-19)	(0.01, 0.4) (0.1, 0.0) (0.01, 0.4)
	Lung Segmentation (NLM)	(0.016, 0.6) (0.016, 0.2) (0.001, 0.6)
	Malaria Cell Detection (NLM dataset sizes)	(0.01, 0.8) (0.016, 0.825) (0.1, 0.6) (0.016, 0.4)
	Malaria Cell Detection (NLM noise levels)	(0.01, 0.8) (0.01, 0.85) (0.1, 0.0) (0.1, 0.4)
Adam	Digit Classification (MNIST noise levels)	(0.001, 0.9) (0.001, 0.0) (0.001, 0.0) (0.001, 0.85) (0.001, 0.85)
	TB Classification (TBX11K dataset sizes)	(0.0001, 0.825) (0.001, 0.85) (0.001, 0.0) (0.001, 0.85)
	Lung Segmentation (COVID-19)	(0.0001, 0.6) (0.0001, 0.6) (0.001, 0.825)
	Lung Segmentation (NLM)	(0.0001, 0.0) (0.0001, 0.874), (0.0001, 0.4)
	Malaria Cell Detection (NLM dataset sizes)	(0.0001, 0.0) (0.001, 0.0) (0.0001, 0.6) (0.0001, 0.0)
	Malaria Cell Detection (NLM noise levels)	(0.0001, 0.0) (0.001, 0.2) (0.001, 0.2) (0.01, 0.925)

Table 1: Summary of learning rates and momentum values for SGD and Adam that provided the best performance for variations of dataset size and noise levels.

noise, a learning rate of (0.1) was preferred, again in combination with smaller momentum values of zero (Poisson noise) or 0.4 (salt-and-pepper noise). Thus, learning rates around 0.1 performed well again and lower momentum values were observed as well.

3.5.2 Adam

For the MNIST dataset, in contrast to SGD, Adam maintained a constant learning rate of 0.001 across all noise levels. However, the momentum varied, starting at 0.9 with no noise, dropping to 0.0 for 5% and 10% noise, and then increasing again to 0.85 for 15% and 20% noise.

In the context of tuberculosis classification using the TBX11K dataset with different training set sizes, the optimal learning rate and momentum combinations were identified as follows: With 100% training data, the optimal pair was (0.0001, 0.825); with 75% of the data, it was (0.001, 0.85); with 50%, the best combination was (0.001, 0.0); and finally with 25% of the data, it was again (0.001, 0.85). Thus, the optimal learning rate was always very small and preferred larger momentum values, with one exception.

For lung segmentation on the COVID-19 and NLM datasets, the best-performing models of Adam were associated with lower learning rates compared to SGD, regardless of momentum. The optimal learning rate stayed consistently at 0.0001, paired with a range of momentum values, varying from none (0.0) to moderate (0.6) and as high as high (0.85).

Similarly, for the malaria dataset and different training set sizes, Adam performed best for a small learning rate of 0.0001 or 0.001 with momentum values of 0.0 and 0.6. Under different noise conditions, learning rates of 0.0001, 0.001, and 0.01, with a wide momentum range of 0.0 to 0.925, were observed.

4. DISCUSSION

The current literature lacked a comprehensive analysis of hyperparameters for optimization strategies such as SGD and Adam. This study presented extensive grid search experiments for both of these strategies. The experiments thoroughly explored a broad range of settings for learning rates and momentum. This extensive exploration was not limited to a single dataset or task; instead, it spanned multiple datasets and encompassed a

variety of tasks to increase the breadth and applicability of the findings. Numerous experiments presented in this paper supported the perspective that SGD offered a superior generalization performance under noise, limited training data, or domain shifts, whereas Adam seemed to be more prone to over-training and memorizing.

In the case of noise, one notable paper in the literature was the work by Subhajit et al., who investigated the training of neural network models on datasets corrupted with noise, employing a range of different optimizers.[25] Their aim was to understand how various optimizers respond to noisy data during training. They evaluated the trained models on clean data to assess and benchmark the extent to which the optimizers might overfit given the noise present in the training set. Their results demonstrated the robustness of SGD compared to adaptive optimization like Adam and RMSProp under noisy training data.

As shown in Table 1, Adam consistently performed best for very small learning rates, which were smaller than the best rates observed for SGD. On the other hand, the momentum appeared to have no significant impact on Adam's performance. For SGD, higher momentum values provided the best performance, starting from 0.4 up to 0.8 and higher. In summary, the experiments corroborated the values suggested for SGD in,[1] namely a learning rate of 0.016 and a momentum value of 0.874. Moreover, the experiments showed that SGD, when optimally tuned, could generalize better than Adam, especially in noisy scenarios or in case of small training data and domain shifts.

5. CONCLUSIONS

The study concluded that biased outcomes in machine learning models, influenced by factors such as limited training data, imbalanced class distribution, and overfitting, can be mitigated by the careful selection of optimizers and hyperparameters in neural network training. Employing SGD and Adam optimizers with various hyperparameters, the study investigated classification, segmentation, and detection tasks across diverse image-based applications. For the MNIST handwritten digit dataset, SGD consistently surpassed Adam, particularly under increasing noise conditions, displaying a superior generalization performance. For tuberculosis detection using DenseNet121, SGD excelled in smaller training sets, while Adam was more effective on larger sets. For lung segmentation with the COVID-19 and NLM chest X-ray datasets, both optimizers showed comparable accuracy, with SGD performing better spatial measures. The study further evaluated the performance of SGD and Adam in identifying malaria parasites. Both optimizers achieved a high accuracy on the full-size datasets, but unlike Adam, SGD maintained a consistently high accuracy with decreasing dataset sizes and in various noise conditions. Additional experiments addressed domain-shift challenges from one country to another, demonstrating that SGD could adapt better to a different data source compared to Adam. This comprehensive investigation highlighted the significance of selecting appropriate optimizers and hyperparameters in enhancing the generalizability of machine learning models, thereby contributing to the development of unbiased and fair machine learning practices.

ACKNOWLEDGMENTS

This work was supported in part by the Lister Hill National Center for Biomedical Communications of the National Library of Medicine (NLM), National Institutes of Health. The work has also been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services under BCBB Support Services Contract HHSN316201300006W/-75N93022F00001 to Guidehouse Inc. This work utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov).

REFERENCES

[1] Jaeger, S., "The golden ratio in machine learning," in [*IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*], 1–7 (2021).

[2] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," in [*arXiv e-prints*], (2014).

[3] Hardt, M., Recht, B., and Singer, Y., "Train faster, generalize better: Stability of stochastic gradient descent," in [*International conference on machine learning*], 1225–1234 (2016).

[4] Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B., "The marginal value of adaptive gradient methods in machine learning," *Advances in neural information processing systems* **30** (2017).
[5] Keskar, N. S. and Socher, R., "Improving generalization performance by switching from Adam to SGD," in [*arXiv e-prints*], (2017).
[6] Chen, J., Zhou, D., Tang, Y., Yang, Z., Cao, Y., and Gu, Q., "Closing the generalization gap of adaptive gradient methods in training deep neural networks," in [*In IJCAI*], (2020).
[7] Zhou, Y., Karimi, B., Yu, J., Xu, Z., and Li, P., "Towards theoretically understanding why SGD generalizes better than Adam in deep learning," *Advances in Neural Information Processing Systems* **33**, 21285–21296 (2020).
[8] Xie, Z., Wang, X., Zhang, H., Sato, I., and Sugiyama, M., "Adaptive inertia: Disentangling the effects of adaptive learning rate and momentum," in [*In International conference on machine learning*], 24430–24459 (2022).
[9] Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J., and Dalca, A. V., "Voxelmorph: a learning framework for deformable medical image registration," *IEEE transactions on medical imaging* **38**, 1788–1800 (2019).
[10] Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., and Rueckert, D., "Self-supervised learning for medical image analysis using image context restoration," *Medical image analysis* **58**, 101539 (2019).
[11] Li, X., Yu, L., Chen, H., Fu, C.-W., Xing, L., and Heng, P.-A., "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems* **32**, 523–534 (2020).
[12] Apostolopoulos, I. D. and Mpesiana, T. A., "Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and engineering sciences in medicine* **43**, 635–640 (2020).
[13] Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., and Yang, B., "Medgan: Medical image translation using gans," *Computerized medical imaging and graphics* **79**, 101684 (2020).
[14] He, K., Zhang, X., Ren, S., and Sun, J., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in [*In Proceedings of the IEEE international conference on computer vision (ICCV)*], 1026–1034 (2015).
[15] Luo, L., Chen, H., Xiao, Y., Zhou, Y., Wang, X., Vardhanabhuti, V., Wu, M., and et al, "Rethinking annotation granularity for overcoming shortcuts in deep learning-based radiograph diagnosis: A multicenter study," *Radiology. Artificial intelligence* **4**, e210299 (2022).
[16] "TBX11K chest x-ray dataset," (2020). https://mmcheng.net/tb/, last accessed July 2023.
[17] Edwardsson, S. and Rizzoli, A., "COVID-19 xray dataset," (2020). https://github.com/v7labs/covid-19-xray-dataset, last accessed September 2023.
[18] Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., and Thoma, G., "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery* **4**, 475–477 (2014).
[19] "Montgomery chest x-ray dataset," (2014). https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Montgomery-County-CXR-Set/MontgomerySet/index.html, last accessed July 2023.
[20] "Shenzhen chest x-ray dataset," (2014). https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Shenzhen-Hospital-CXR-Set/index.html, last accessed July 2023.
[21] "Ultralytics YOLOv8," (2023). https://github.com/ultralytics/ultralytics, last accessed September 2023.
[22] Silamut, M. P. K., Maude, R. J., Jaeger, S., and Thoma, G., "Image analysis and machine learning for detecting malaria," *Translational research: the journal of laboratory and clinical medicine* **194**, 36–55 (2018).
[23] "NLM malaria dataset," (2018). https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html, last accessed July 2023.

[24] Yu, H., Mohammed, F. O., Hamid, M. A., Yang, F., Kassim, Y. M., Mohamed, A. O., Maude, R. J., and et al., "Patient-level performance evaluation of a smartphone-based malaria diagnostic application," *Malaria Journal* **22**, 1–105 (2023).

[25] Chaudhury, S. and Yamasaki, T., "Robustness of adaptive neural network optimization under training noise," *IEEE Access* **9**, 37039–37053 (2021).