Inflation of test accuracy due to data leakage in deep learning-based classification of OCT images

Iulian Emil Tampu[1,2,*], **Anders Eklund**[1,2,3], **and Neda Haj-Hosseini**[1,2]

[1]Department of Biomedical Engineering , Linköping University, 581 85 Linköping, Sweden
[2]Center for Medical Image Science and Visualization , Linköping University, 581 83 Linköping, Sweden
[3]Division of Statistics & Machine Learning, Department of Computer and Information Science, Linköping University, 581 83 Linköping, Sweden
[*]corresponding author: Iulian Emil Tampu (iulian.emil.tampu@liu.se)

ABSTRACT

In the application of deep learning on optical coherence tomography (OCT) data, it is common to train classification networks using 2D images originating from volumetric data. Given the micrometer resolution of OCT systems, consecutive images are often very similar in both visible structures and noise. Thus, an inappropriate data split can result in overlap between the training and testing sets, with a large portion of the literature overlooking this aspect. In this study, the effect of improper dataset splitting on model evaluation is demonstrated for three classification tasks using three OCT open-access datasets extensively used in the literature, Kermany's and Srinivas ophthalmology datasets ,and AIIMS breast tissue dataset. Our results show that the classification performance is inflated by 7.5 up to 21.7 percentage units in terms of Matthew correlation coefficient for models tested on a dataset with improper splitting, highlighting the considerable effect of dataset handling on model evaluation. This study intends to raise awareness on the importance of dataset splitting given the increased research interest on implementing deep learning on OCT data.

Introduction

The evaluation of deep learning models, and in general machine learning methods, aims at providing an unbiased description of model performance. Given a pool of data suitable for studying a hypothesis (e.g. classification, regression or segmentation), different splits of the data are commonly created for model training, model hyper-parameter tuning (validation set) and model assessment (testing set). This translates to having a part of the data used to fit model parameters and tune model hyperparameters, and another set to assess model generalization on unseen data[1]. Disregarding the choice of having separate validation and testing splits[2], the strategy used to generate the testing set from the original pool of data can have a large impact on the assessment of the model's final performance. Several studies have investigated the effect of the relative size between training, validation and/or testing sets[1,3] on model performance, as well as how training and validation sets can be used via resampling techniques, such as cross-validation, during model training[1,4]. More importantly, it is widely accepted that no overlap should exist between the samples used for model fitting and hyperparameter tuning, and those belonging to the testing set. If overlap is present, the model performance will be biased and uninformative with respect to the generalization capabilities of the model to new samples. However, although trivial, when implementing data splitting strategies the overlap between training and testing sets can be easily overlooked. This is especially true for 3D medical image data, as 2D methods are often applied to individual slices given the limited amount of memory of graphics hardware. A proper splitting must therefore be done on the volume or subject level and not on slice level.

Nowadays, machine learning methods, especially convolutional neural networks (CNNs) and deep learning algorithms, are widely used in research to analyze medical image data. A plethora of publications describe CNN implementations on a variety of both 2D and 3D medical data[5–7]. Reliable evaluation of such methods is paramount since this informs the research community on the models' performance, allows meaningful comparison between methods, and to a greater extent indicates which research questions might be worth further investigation. To this regard, many medical image analysis challenges were established that aim at providing an unbiased platform for the evaluation and ranking of different methods on a common and standard testing dataset. In a review, Maier-Hein et al.[8] have collected the majority of the available medical image competitions and analyzed the reliability of such challenges. Despite the many limitations discovered, including missing information regarding how the ground truth was obtained, the authors emphasized the importance of these challenges and their contribution towards a more transparent and reliable evaluation of deep learning methods for medical image applications.

However, not all implementations can be evaluated through dedicated challenges. Thus, when such third-party eval-

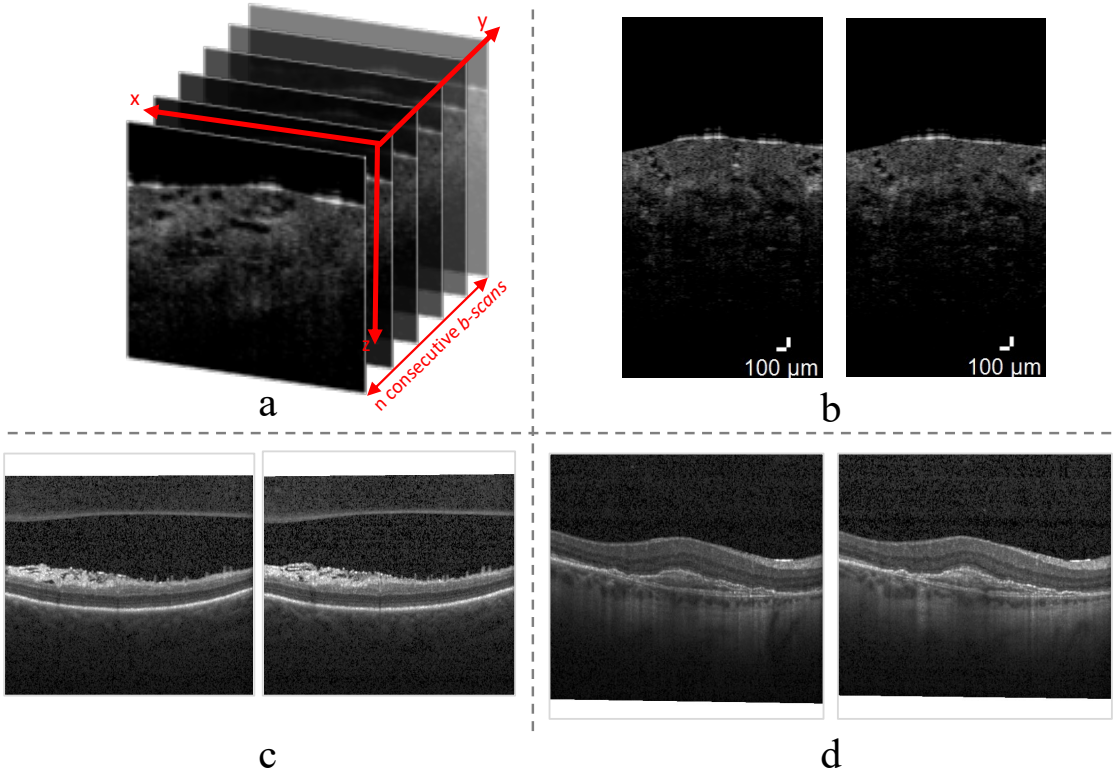

Figure 1. Schematic of an OCT volume with examples of consecutive slices (*b-scans*) from the three open-access OCT datasets used in this study. (a) pictures the consecutive 2D *b-scans* rendering a 3D OCT volume. Here an example from the AIIMS dataset[14] is used for illustrative purposes. In (b) the consecutive *b-scans* separated by ∼18 micrometers are examples of healthy breast tissue (Patient 15, volume 0046, slices 0075 and 0076) from the AIIMS dataset[14]. (c) shows consecutive images of retina affected by choroidal neovascularization (CNV 81630-33 and 81630-34) from the Kermany's OCT2017 dataset[15]. (d) images of age-related macular degeneration (AMD) from Srinivas dataset (AMD2 046 and 047). Note that the *b-scans* from both Kermany's and Srinivas' dataset are given with data augmentation applied.

uation platforms are not available, it is the responsibility of the researchers performing the investigation to evaluate their method thoroughly. As for the case of many of the reviewed medical image analysis challenges[8], one aspect that is sometimes missing or not well described is how the testing dataset is generated from the original pool of data. Moreover, there are also examples where the preparation of the testing dataset is described, but its overlap with the training set was not considered[9–13], undermining the reproducibility of the study as well as the reliability of the reported results. This is specifically more common in those applications where, due to hardware limitations or model design choices, the data from one subject (or acquisition) is used to obtain multiple dimensionally-smaller samples used for model training and testing. An example of such a scenario is the slicing of a 3D volume into 2D images. In these cases, the overlap between training and testing sets results from having 2D images from the same subject (or acquisition) belonging to both sets.

Focusing on deep-learning applications for optical coherence tomography (OCT), the pool of data used for model training and testing commonly originates from volumetric acquisitions of the same samples under investigation. Depending on the acquisition set-up, volumes are usually acquired with micrometer resolution in the x, y and z directions in a restricted field of view, with tissue structures that are alike and affected by similar noise. This results in consecutive slices having a high similarity., both structure and noise. Figure 1 shows a schematic of an OCT volume along with examples of two consecutive slices from OCT volumes from three open-access datasets[14,15,24].

The majority of the reviewed literature implementing deep learning on OCT data used 2D images from scanned volumes, where two methods were commonly used to split the pool of image data into training and testing sets: *per-image* or *per-volume/subject* splitting. In the *per-volume/subject* splitting approach, the random selection of data for the testing set is done on the volumes (or subjects) ensuring that images from one volume (or subject) belong to either one of the training or testing sets. It is important to notice that even a *per-volume* split might not be enough to avoid overlap between the training and testing

set. In fact, if multiple volumes are acquired from the same tissue region, the tissue structures will be very similar among the volumes. In these scenarios, a *per-subject* split is more appropriate. Overall, assuming that volumes are acquired from different tissue regions, splitting the dataset *per-volume* or *per-subject* (here called *per-volume/subject*) ensures that overlap between training and testing is not present. On the other hand in the *per-image* approach, 2D images belonging to the same volume are considered independent. Thus, the testing set is created by random selection from the pool of images without considering that images from the same volume (or subject) might end up in both the training and testing sets. Even if this approach clearly results in overlap between the the testing and training sets, a number of reviewed studies as well as one of the most downloaded open-access OCT dataset employed a *per-image* split.

Thus, the aim of this study is to demonstrate the effect of improper dataset splitting (*per-image*) on classification accuracy using three open-access OCT datasets, Kermany's OCT2017[15], AIIMS[14] and Srinivas[24]. These were selected among the other open-access datasets for several reasons: (1) they are examples of OCT medical images belonging to different medical disciplines (ophthalmology and breast oncology) and showing different tissue structures and textures, (2) they are used in literature to evaluate deep learning-based classification of OCT images, with Kermany's dataset[15] extensively used for developing deep learning methods in ophthalmology (over 14,500 downloads)[16], and (3) the datasets are provided in two different ways, per-subject for the AIIMS and Srinivas dataset, and already split for Kermany's datasets. The latter is an important aspect to consider since many of the studies using Kermany's dataset overlooked the overlap between the training and the testing data.

Results

LightOCT model classification performance on the three datasets and for the different dataset split strategies, is summarized in Table 1, with results presented as mean±standard deviation (mean±std) over the ten times repeated five-fold cross validation. In addition, Figure 2a shows in details the MCC distribution as box plots. For all datasets, the *per-image* split strategy results in a higher model performance compared to the *per-volume/subject* strategy. In particular, looking at the results on Kermany's dataset, the mean MCC dropped by 7.9 percentage units when shifting from a *per-image* strategy to a *per-volume/subject* one. An even larger decrease in performance is seen when comparing the model trained on the *per-volume/subject* with the one trained on the *original split*, which contains overlapping testing and training images, with mean MCC and AUC reduced by 14.2 and 7 percentage units, respectively. A similar trend can be seen for the both the AIIMS and Srinivas datasets, where the model trained on the dataset using *per-image* strategy had a mean MCC higher by 7.5 and 21.7 percentage units, respectively, compared to the one trained on a *per-volume/subject* split.
Results on the random label experiment are presented in Figure 2b for all datasets and dataset split strategies. These show that the mean MCC is close to zero for all the tests performed, describing the random classification of the model.

Discussion

Dataset split should be carefully designed to avoid overlap between training and testing sets. Table 2 summarizes several studies on deep learning applications for OCT data, specifying the described data split strategy. All the works using the *original split* from Kermany's dataset or a *per-image* split strategy reported accuracies >95%. Obtained results are in accordance with the reported values on Kermany's original split, where the difference in performance can be attributed to the optimization of the LightOCT model. Interestingly, Table 2 also shows that studies using multiple datasets and reporting different split strategies for each dataset[10,17,18], show as high accuracy on the *per-volume/subject* split datasets as the one on the *original split* or the *per-image* split datasets. In the light of our results and the overlap found between the training and testing sets in Kermany's *original* split, it is reasonable to ask whether the high performances reported for the *per-volume/subject* split datasets reflect the true high performance of the implemented methods, or are examples of inflated accuracy values due to data leakage between training and testing sets.

Another source of data leakage, which was not investigated in this study but that could similarly inflate classification performance, is data augmentation. In particular, an image could be augmented many time with its different augmented versions ending in the training and testing set. This result in a data overlap similar to the one of a *per-image* split strategy, where images with the same structures and noise properties are in both training and testing sets. The *original* split in Kermany dataset is provided with already augmented images., however, overlap between training and testing with respect to data augmentation was not checked.

In conclusion, the used dataset split strategy can have a substantial impact on the evaluation of deep learning models. In this paper it is demonstrated that, in OCT image classification applications specifically, a *per-image* split strategy of the volumetric data adopted by a considerable number of studies, returns over-optimistic results on model performance and an inflation of test accuracy values, questioning the reliability of the assessments and hindering an objective comparison between research outcomes. This problem has also been demonstrated in 3D magnetic resonance imaging studies[13] and in digital

Table 1. LightOCT model performance on Kermany's[15], AIIMS[14] and Srinivas[24] datasets with training, validation and testing sets split using different strategies. Performance metrics are reported as mean±standard deviation (mean±std) over the models trained through the ten times repeated five-fold cross validation and classes.

Dataset	*Split strategy*	*MCC* [-1,1] *(mean±std)*	*AUC* [0,1] *(mean±std)*	*F1-score* [0,1] *(mean±std)*	*Accuracy* [0,1] *(mean±std)*	*Precision* [0,1] *(mean±std)*	*Recall* [0,1] *(mean±std)*
Kermany's[15] dataset	Results from[10] using *original split*	/	/	/	∼0.96	/	0.945
	original split	0,906	0,992	0,927	0,928	0,979	0,928
	per-image split	0,702±0,015	0,953±0,003	0,758±0,015	0,765±0,013	0,891±0,006	0,765±0,013
	per-volume/subject split	0,623±0,015	0,915±0,006	0,681±0,018	0,694±0,014	0,804±0,011	0,694±0,014
AIIMS[14] dataset	Results from[10] using *per-image split*	/	0.996	0.989	0.988	0.993	0.986
	per-image split	0,982±0,006	0,999±0,001	0,991±0,003	0,9911±0,003	0,999±0,001	0,991±0,003
	per-volume/subject split	0,907±0,106	0,996±0,007	0,949±0,064	0,950±0,061	0,995±0,009	0,950±0,061
Srinivas[24] dataset	Results from[10] using *per-image split*	/	/	/	>0.96	/	0. 988
	per-image split	0,881±0,016	0,987±0,003	0,920±0,011	0,920±0,011	0,976±0,006	0,920±0,011
	per-volume/subject split	0,447±0,109	0,828±0,054	0,607±0,0712	0,617±0,071	0,720±0,079	0,617±0,071

pathology[19], where data leakage between the training and testing sets resulted in over-optimistic classification accuracy (>29% slide level classification accuracy in MR studies and up to 41% in digital pathology).Moreover, greater attention should be paid to the structure of datasets made available to the research community to avoid biasing the evaluation of different methods and undermining the usefulness of open-access datasets. With the increase interest of research community in the implementation of deep learning methods for OCT image analysis, the presented results want to raise awareness on a trivial but overlooked problem that can spoil research efforts.

Methods

Datasets description

The first dataset used in this study is Kermany's ophthalmology dataset[15,35], which is used by an extensive number of other studies (see Table 2). This dataset contains 84,484 2D images of healthy retina (n=26,565) as well as retina affected by Choroidal Neovascularization (CNV) (n=37,455), Diabetic Macular Edema (DME) (n=11,598) and Drusen (n=8,866) from 4686 patients. The images are given as TIFF files of size 512×496 pixels saved after data augmentation (rotation and horizontal flip) and arranged in training and testing sets, with 1000 images from 633 patients, *i.e.*, 250 images from each class set for testing and the remaining for training. In this paper, the third version of the dataset available at[15] was used and hereafter referred to as *original* split. The difference between the versions is in the way the dataset can be downloaded. The authors of the dataset[15] do not specify if the split of the dataset in training and testing sets was performed *per-image* or *per-volume/subject*. By performing an automatic check on the original splits (assuming that the naming convention is CLASS_subject-ID_bscan-ID), it was found that 92% of the test images belong to subject-IDs also found in the training set. Moreover, by visually inspecting the given splits it was possible to identify images in the testing set that were similar to the training set (an example of such a case is training image=DRUSEN-8086850-6, testing image=DRUSEN-8086850-1).

The second used in this study is Srinivas ophthalmology dataset[24] collecting a total of 3,231 2D OCT images of age-related macular degeneration (AMD), diabetic macular edema (DME) and normal subjects. For every class data from 15 subjects was provided as independent folders. OCT images are given as TIFF files with image 512×496 pixels saved after data augmentation (rotation and horizontal flip).

The third and last dataset used in this study is the AIIMS dataset, which is a collection of 18,480 2D OCT images of healthy (n=9,450) and cancerous (n=9,030) breast tissue[14]. The images are obtained from volumetric acquisitions and are provided as BMP files of size 245×442 pixels organized per-class and per-subject (22 cancer subjects and 23 healthy subjects).

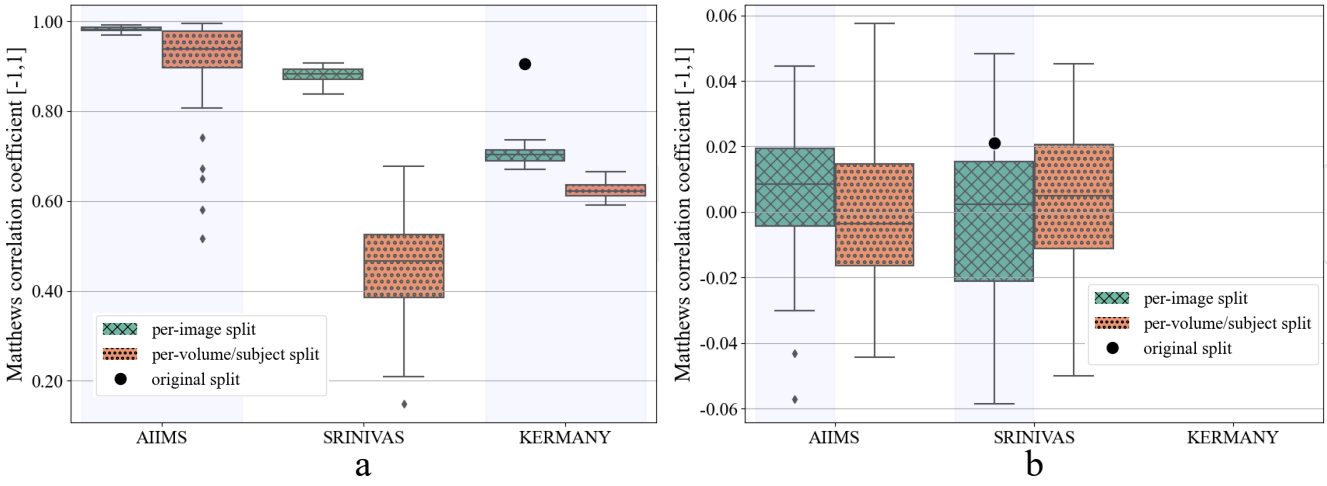

Figure 2. Comparison between Matthews correlation coefficient comparison for LightOCT model trained on different dataset split strategies. In (a) the results with each sample having the correct label, whereas in (b) the results from the random label experiment. For both (a) and (b) each box plot summarise the test MCC for the 50 models trained through a ten times repeated five-fold cross validation. Results are presented for all the three datasets with the *per-image* split strategy shown in striped-green and *per-volume/subject* split strategy in dotted-orange. For Kermany's dataset, the result of the models trained on the *original* split is shown as full-black circle. Figure b needs to include the results from Kermany's trainings

For all datasets, a custom split function was implemented to split the dataset *per-image* or *per-volume*. In either case, 1,000 images from every class were assigned for testing. In the case of Kermany's dataset, only one of the 9 reviewed studies using Kermany's dataset reported on the overlap between the training and testing splits, and re-split the data using a *per-volume/subject* strategy[30]. All the other studies (see Table 2) used the *original split*, biasing the evaluation of their implemented models. Example images of these three datasets are shown in Figures 1b, 1c and 1d.

Model architecture and training strategy

The LightOCT model proposed by Butola et al.[10] was used in this study. LightOCT is a custom, shallow and multi-purpose network for OCT image classification composed of a two-layer CNN encoder, and one fully connected layer with softmax activation as output layer. The first and second convolutional layers have 8 and 32 convolutional filters, respectively. The kernel size of the filters in both layers is set to 5×5 and the output of each layer passes through a ReLU activation function[10]. A max-pooling operation is present between the first and the second convolutional layer that reduces in half the spatial dimension of the output of the first layer. The two-dimensional output of the CNN encoder is then flattened to a one dimensional vector which is fed to the fully connected layer for classification. The number of nodes in the fully connected layer is changed based on the number of classes specified by the classification task[10].

For all of the classification tasks, the model was trained from scratch using stochastic gradient descent with momentum (m=0.9) with constant learning rate (lr=0.0001). For all experiments, the batch size was set to 64 and the model was trained for 250 epochs without early stopping. Note that model architecture and training hyperparameters were not optimized for each dataset since it was out of the scope of this work. The model architecture as well as the training hyperparameters were chosen based on the results of Butola et al.,[10]. The model and the training routine were implemented in Tensorflow 2.6.2, and training was run on a computer with 20 core CPU and 4 Nvidia Tesla V100 GPUs.

Evaluation metrics

Models were trained on the *original* split, if available, and on training and testing splits obtained using a *per-image* and *per-volume/subject* strategy. A ten times repeated five-fold cross validation was run for both split strategies to ensure reliability of the presented results. A multi-class confusion matrix was used to evaluate the classification performance of the model with Matthews correlation coefficient (MCC) obtained as a derived metrics coherent with respect to class imbalance and stable to label randomization?. Accuracy, precision, recall and F1-score computed were also derived for each class using the definitions provided by Sokolova et al.,[36] to allow comparison with previous studies. Additionally, receiver operator characteristic (ROC) curves were used along with the respective area under the curve (AUC). A random label experiment was also carried out where the models were trained and tested on samples with randomized labels. The scope of the random label experiment was to

Table 2. Summary of reviewed literature with focus on dataset split and reported test classification performance. Open-access datasets and the ones available upon request are marked by * and **, respectively. Dataset is not open-access if not specified. Datasets obtained from animal model samples are marked by †. The difference in performance between studies using the same datasets results from the different implemented methods.

Ref.	*OCT dataset*	*Data split strategy*	*Model performance on testing set*
9	Data from thyroid, parathyroid, fat and muscle samples	*per-image*	97.12% accuracy
20	Ophthalmology[15]*	*original split*	95.55% accuracy
21	Ophthalmology[15]*	*original split*	99.1% accuracy
22	Ophthalmology[15]*	*original split*	98.7% accuracy
22	Ophthalmology[15]*	*original split*	96.6% accuracy
23	Ophthalmology[15]*	*original split*	99.6% accuracy
17	(1) Ophthalmology[15]* (2) Ophthalmology[24]*	(1) *original split* (2) *per-volume/subject*	(1) 99.80% accuracy (2) 100% accuracy
25	Coronary artery OCT	*per-volume/subject*	96.05% accuracy
26	Kidney†	*per-volume/subject*	82.6% accuracy
27	Data from high and low grade brain tumors	*per-volume/subject*	97% accuracy
28	Colon**,†	*per-volume/subject*	88.95% accuracy on 2D images
29	Data from breast tissue	*per-volume/subject*	91.7% specificity
30	Ophthalmology[15]*	*per-volume/subject*	98.46% accuracy
10	(1) Ophthalmology[15]* (2) Ophthalmology[24]* (3) Breast tissue[14]*	(1) *original split* (2) *per-volume/subject* (3) *per-image*	(1) 96% accuracy (2) > 98.8% accuracy (3) 98.8% accuracy
18	(1) Ophthalmology[24]* (2) Ophthalmology[31]* (3) Ophthalmology[32]* (4) Ophthalmology[15]*	(1) *per-volume/subject* (2) *per-volume/subject* (3) *per-volume/subject* (4) *original split*	(1) 96.66% accuracy (2) 98.97% accuracy (3) 99.74% accuracy (4) 99.78% accuracy
33	Dentistry	No description given	98% sensitivity 100% specificity
34	Ophthalmology	No description given	99.19% accuracy

highlight any bias in the data or training, which would result in a MCC different from zero.

Acknowledgments

The study was supported by the grants from Åke Wiberg Stiftelse (M19-0455, M20-0034, M21-0083), FORSS - 931466, Vinnova project 2017-02447 via Medtech4Health and Analytic Imaging Diagnostics Arena (1908) and Swedish research council (2018-05250).

Dataset availability statement

The datasets used in this study are open-access, with the AIIMS dataset[14] available at https://www.bioailab.org/datasets, Kermany's OCT2017[15] at https://data.mendeley.com/datasets/rscbjbr9sj/3. and Srinivas at https://people.duke.edu/~sf59/Srinivasan_BOE_2014_dataset.htm.

Code availability

The code used to generate the results in this paper is available at https://github.com/IulianEmilTampu/OCT_SPLIT_PROPERLY_YOUR_DATA.git

Author contributions statement

IET contributed with conceptualization, methodology, code development and implementation, formal analysis and drafting the manuscript. AE contributed with supervision and hardware resources. NHH contributed with conceptualization, supervision and funding. All authors contributed to the interpretation of the results, have revised and edited the manuscript and approved the submitted version.

Competing interests

AE has previously received Nvidia hardware for research.

References

1. Xu, Y. & Goodacre, R. On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning. *J. analysis testing* **2**, 249–262 (2018).
2. Kuhn, M., Johnson, K. *et al. Applied predictive modeling*, vol. 26 (Springer, 2013).
3. Guyon, I. *et al.* A scaling law for the validation-set training-set size ratio. *AT&T Bell Lab.* **1** (1997).
4. Refaeilzadeh, P., Tang, L. & Liu, H. Cross-validation. *Encycl. database systems* **5**, 532–538 (2009).
5. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. image analysis* **42**, 60–88 (2017).
6. Ker, J., Wang, L., Rao, J. & Lim, T. Deep Learning Applications in Medical Image Analysis. *Ieee Access* **6**, 9375–9389 (2017).
7. Anwar, S. M. *et al.* Medical Image Analysis using Convolutional Neural Networks: A Review. *J. medical systems* **42**, 1–13 (2018).
8. Maier-Hein, L. *et al.* Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. communications* **9**, 1–13 (2018).
9. Wang, H., Won, D. & Yoon, S. W. A deep separable neural network for human tissue identification in three-dimensional optical coherence tomography images. *IISE Transactions on Healthc. Syst. Eng.* **9**, 250–271 (2019).
10. Butola, A. *et al.* Deep learning architecture "LightOCT" for diagnostic decision support using optical coherence tomography images of biological samples. *Biomed. Opt. Express* **11**, 5017–5031 (2020).
11. Irmak, E. Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework. *Iran. J. Sci. Technol. Transactions Electr. Eng.* **45**, 1015–1036 (2021).
12. Sadad, T. *et al.* Brain tumor detection and multi-classification using advanced deep learning techniques. *Microsc. Res. Tech.* **84**, 1296–1308 (2021).
13. Yagis, E. *et al.* Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Sci. reports* **11**, 1–13 (2021).
14. Butola, A. *et al.* Volumetric analysis of breast cancer tissues using machine learning and swept-source optical coherence tomography. *Appl. optics* **58**, A135–A141 (2019).
15. Kermany, D., Zhang, K. & Goldbaum, M. Large Dataset of Labeled Optical Coherence tomography (OCT) and Chest X-Ray images. *Mendeley Data* **3**, 10–17632 (2018).
16. Retinal OCT Images (optical coherence tomography). https://kaggle.com/paultimothymooney/kermany2018. Accessed: 2022-02-10.
17. Kamran, S. A., Saha, S., Sabbir, A. S. & Tavakkoli, A. Optic-Net: A Novel Convolutional Neural Network for Diagnosis of Retinal Diseases from Optical Tomography Images. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, 964–971 (IEEE, 2019).
18. Thomas, A. *et al.* A novel multiscale and multipath convolutional neural network based age-related macular degeneration detection using OCT images. *Comput. Methods Programs Biomed.* **209**, 106294 (2021).

19. Bussola, N., Marcolini, A., Maggio, V., Jurman, G. & Furlanello, C. AI slipping on tiles: Data leakage in digital pathology. In *International Conference on Pattern Recognition*, 167–182 (Springer, 2021).

20. Najeeb, S. *et al.* Classification of retinal diseases from OCT scans using convolutional neural networks. In *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*, 465–468 (IEEE, 2018).

21. Chen, Y.-M., Huang, W.-T., Ho, W.-H. & Tsai, J.-T. Classification of age-related macular degeneration using convolutional-neural-network-based transfer learning. *BMC bioinformatics* **22**, 1–16 (2021).

22. Latha, V., Ashok, L. & Sreeni, K. Automated Macular Disease Detection using Retinal Optical Coherence Tomography images by Fusion of Deep Learning Networks. In *2021 National Conference on Communications (NCC)*, 1–6 (IEEE, 2021).

23. Tsuji, T. *et al.* Classification of optical coherence tomography images using a capsule network. *BMC ophthalmology* **20**, 1–9 (2020).

24. Srinivasan, P. P. *et al.* Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed. optics express* **5**, 3568–3577 (2014).

25. Athanasiou, L. S., Olender, M. L., José, M., Ben-Assa, E. & Edelman, E. R. A deep learning approach to classify atherosclerosis using intracoronary optical coherence tomography. In *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950, 163–170 (SPIE, 2019).

26. Wang, C. *et al.* Deep-learning-aided forward optical coherence tomography endoscope for percutaneous nephrostomy guidance. *Biomed. optics express* **12**, 2404–2418 (2021).

27. Gesperger, J. *et al.* Improved diagnostic imaging of brain tumors by multimodal microscopy and deep learning. *Cancers* **12**, 1806 (2020).

28. Saratxaga, C. L. *et al.* Characterization of Optical Coherence Tomography Images for Colon Lesion Differentiation under Deep Learning. *Appl. Sci.* **11**, 3119 (2021).

29. Singla, N., Dubey, K. & Srivastava, V. Automated assessment of breast cancer margin in optical coherence tomography images via pretrained convolutional neural network. *J. biophotonics* **12**, e201800255 (2019).

30. Chetoui, M. & Akhloufi, M. A. Deep retinal diseases detection and explainability using OCT images. In *International Conference on Image Analysis and Recognition*, 358–366 (Springer, 2020).

31. Rasti, R., Rabbani, H., Mehridehnavi, A. & Hajizadeh, F. Macular OCT classification using a multi-scale convolutional neural network ensemble. *IEEE transactions on medical imaging* **37**, 1024–1034 (2017).

32. Farsiu, S. *et al.* Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology* **121**, 162–172 (2014).

33. Karimian, N., Salehi, H. S., Mahdian, M., Alnajjar, H. & Tadinada, A. Deep learning classifier with optical coherence tomography images for early dental caries detection. In *Lasers in Dentistry XXIV*, vol. 10473, 1047304 (International Society for Optics and Photonics, 2018).

34. Wang, R. *et al.* OCT image quality evaluation based on deep and shallow features fusion network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 1561–1564 (IEEE, 2020).

35. Kermany, D. S. *et al.* Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).

36. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. processing & management* **45**, 427–437 (2009).