tvst

Article

Learning to Discover Explainable Clinical Features With Minimum Supervision

Lutfiah Al Turk[1,*], Darina Georgieva[2,*], Hassan Alsawadi[4], Su Wang[2], Paul Krause[2], Hend Alsawadi[3], Abdulrahman Zaid Alshamrani[5], George M. Saleh[6], and Hongying Lilian Tang[2]

[1] Department of Statistics, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia
[2] Department of Computer Science, University of Surrey, Guildford, Surrey, UK
[3] Faculty of Medicine, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia
[4] Department of Electrical and Computer Engineering, King Abdulaziz, University, Jeddah, Kingdom of Saudi Arabia
[5] Ophthalmology Department, Faculty of Medicine, University of Jeddah, Kingdom of Saudi Arabia
[6] NIHR Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology, UK

Correspondence: Darina Georgieva, Department of Computer Science, University of Surrey, Guildford, Surrey, UK. e-mail: darina.georgiewa@gmail.com

Received: February 24, 2021
Accepted: November 15, 2021
Published: January 11, 2022

Keywords: semisupervised learning; explainability; transfer learning; OCT; ophthalmology; interpretability

Citation: Al Turk L, Georgieva D, Alsawadi H, Wang S, Krause P, Alsawadi H, Alshamrani AZ, Saleh GM, Tang HL. Learning to discover explainable clinical features with minimum supervision. Transl Vis Sci Technol. 2022;11(1):11, https://doi.org/10.1167/tvst.11.1.11

Purpose: To compare supervised transfer learning to semisupervised learning for their ability to learn in-depth knowledge with limited data in the optical coherence tomography (OCT) domain.

Methods: Transfer learning with EfficientNet-B4 and semisupervised learning with SimCLR are used in this work. The largest public OCT dataset, consisting of 108,312 images and four categories (choroidal neovascularization, diabetic macular edema, drusen, and normal) is used. In addition, two smaller datasets are constructed, containing 31,200 images for the limited version and 4000 for the mini version of the dataset. To illustrate the effectiveness of the developed models, local interpretable model-agnostic explanations and class activation maps are used as explainability techniques.

Results: The proposed transfer learning approach using the EfficientNet-B4 model trained on the limited dataset achieves an accuracy of 0.976 (95% confidence interval [CI], 0.963, 0.983), sensitivity of 0.973 and specificity of 0.991. The semisupervised based solution with SimCLR using 10% labeled data and the limited dataset performs with an accuracy of 0.946 (95% CI, 0.932, 0.960), sensitivity of 0.941, and specificity of 0.983.

Conclusions: Semisupervised learning has a huge potential for datasets that contain both labeled and unlabeled inputs, generally, with a significantly smaller number of labeled samples. The semisupervised based solution provided with merely 10% labeled data achieves very similar performance to the supervised transfer learning that uses 100% labeled samples.

Translational Relevance: Semisupervised learning enables building performant models while requiring less expertise effort and time by using to good advantage the abundant amount of available unlabeled data along with the labeled samples.

Introduction

Each year more than 30 million optical coherence tomography (OCT) scans are obtained to support eye care professionals to identify sight-threatening diseases.[1] OCT is a reliable way to detect eye-related disorders at an early treatable phase and prevent vision loss.[2] The three-dimensional (3D) scans require analysis and interpretation by an expert, but the volume of scans produced is much higher than the number of available healthcare professionals. Consequently, many patients experience long delays until appointed treatment, which in some cases lead to an irreversible visual impairment that could have been prevented by prompt intervention and suitable medical care.

Medical domains are often characterized by an insufficient amount of data samples, which may lead

to the poor ability of a neural network to generalize. Manual labeling imaging data requires professional expertise and is an extremely costly, labor-expensive, and time-consuming process. Generally, medical datasets are composed of more unlabeled data samples compared to labeled ones. Meanwhile, machine learning algorithms are referred to as "black boxes" because of the lack of explainability behind their decisions, which is a prime obstacle to be accepted by the general public and used in critical fields such as medical diagnosis. The medical domain demands much higher than other domains for fidelity, performance and explainability of the neural networks models.

A well-known technique to overcome the challenge of insufficient amount of training data is transfer learning. Transfer learning is a method to leverage knowledge from one domain to another by utilizing the developed abilities of the neural network to recognize generic features such as edges, corners, colors. It enables the training with fairly limited training examples and reaches competitive performance results. On the other side, semisupervised learning techniques are beneficial when the dataset contains both labeled and unlabeled inputs, generally, with a significantly smaller number of labeled samples. Recent successes of semisupervised learning have been the main motivation to investigate its use in the ophthalmological diagnostic domain and compare its performance to the supervised transfer learning-based approach.

Semisupervised learning has started gaining popularity in recent years. The application of the semisupervised learning approach is explored in a limited number of different tasks and domains.[3] In this work, the main focus is medical image analysis and the classification task of OCT data in particular. Previously, semisupervised learning usage for image classification has been explored in medical domains such as brain tumor,[4] breast cancer,[5] chess X-ray,[6] however, no investigation of the applications in the OCT domain has been done to date.

In this work, we designed a transfer learning–based solution using one of the current state-of-the-art neural network architectures, particularly EfficientNet-B4.[7] To further address the problem of insufficient data, we explore the application of semisupervised learning by applying the recently proposed semisupervised framework, SimCLR.[8] Semisupervised learning-based solutions have enormous potential in the medical domain because of the general lack of labeled data in the field. We aim to demonstrate the power of using semisupervised learning by comparing the performance between the supervised transfer learning approach and the semisupervised learning-based solution.

Methods

Dataset Characteristics and Data Splitting

In this project, we use the largest publicly available OCT dataset that is composed of four categories.[9] The dataset contains three groups of retinal pathologies: choroidal neovascularization (CNV), diabetic macular edema (DME), and drusen and one group of healthy samples labeled as normal. A sample representation of all four classes and their representative characteristics are shown in Figure 1.

Precise data splitting is of high importance because each individual patient is represented by multiple scans in the dataset and the OCT scans that belong to the same patient are very similar and often nearly identical. Therefore, the placement of the individual patient's scans in different data partitions such as train, test, validation, might introduce bias to the model and produce misleading performance results. The dataset is highly imbalanced and includes in total 108,312 training images (37,206 CNV, 11,349 DME, 8,617 DRUSEN, and 51,140 NORMAL) from 4,686 patients and 1,000 testing images (250 for each category) from 633 patients. In this study, the dataset is further analyzed and scans that belong to the same patient and are placed in both the train and test partition are eliminated from the training samples. This resulted in reducing the training data from 108,312 to 104,649 with 1,000 set aside for the validation partition. The aim was to keep the testing data the same as provided by the dataset authors while ensuring that the data splits are composed of independent patients' scans which are not included in more than one partition.

A limited dataset is constructed by applying an undersampling of the minority class technique. The limited dataset consists of 7800 samples per training category. The dataset is used to conduct experiments

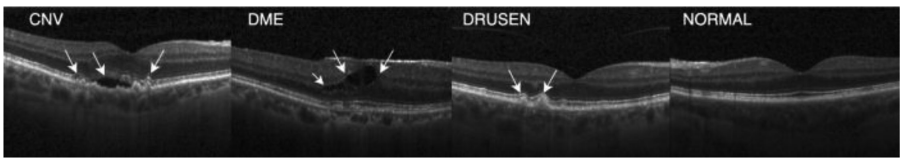

Figure 1. OCT categories: CNV, DME, drusen, and normal.

Table 1. Number of OCT Images per Dataset Partition for the Three Dataset Varieties: Original, Limited and Mini

Dataset	Train	Val	Test	Total
Original	104,649	1000	1000	106,649
Limited	31,200	1000	1000	33,200
Mini	4,000	1000	1000	6,000

Test dataset is the same for all three datasets varieties. Validation dataset is the same for all three datasets varieties.

with fewer data samples to evaluate the performance of the model with smaller amounts of data.

A mini dataset is created and is composed of 1000 image samples per training category. The prime purpose of the mini dataset is to assess the generalization abilities of the neural networks on a very small number of samples.

Table 1 summarizes the used dataset varieties. For a fair comparison, the three datasets—original, limited, and mini—share the same testing and validation datasets.

Data Preprocessing

The data preprocessing flow is illustrated in Figure 2. The size of the images in the dataset varies in width and height;00 therefore the images are resized accordingly to the model requirements (224 × 224 input size). After that, the normalization technique is applied to convert the image pixel values in the range [0, 1], subtract the mean from each image and divide by the standard deviation. To increase the diversity of the data two main data augmentation strategies are applied. Horizontal random flip is one of the strategies deployed, and although it is a very simple method, it is particularly beneficial to the retinal image domain due to the bilateral symmetry of the eye. The technique imitates the inclusion of both left and right eyes in the dataset since the results are biologically feasible. The second data augmentation technique that is implemented is the color space transformation and particularly the color jitter technique of arbitrarily changing the saturation, brightness, and contrast of the image.

Supervised Transfer Learning-Based Solution Details

In deep learning, it is a general perception that to achieve high generalization ability of the model, a large amount of labeled data is required. Building a neural network with data-driven features such as convolutional kernels with an insufficient amount of data can easily result in bad generalization and consequently a negative effect on the evaluation results. Moreover, traditional learning is often associated with long training time and long hyperparameters tuning. However, it still does not guarantee convergence, and it is highly prone to overfitting or underfitting if not an adequate quantity of data is provided. It is particularly challenging to collect a sufficient amount of labeled and verified data in the medical domain because it requires medical professionals to be involved in the process. In addition, patients' consents need to be obtained so that their scans and medical information can be used.

A common strategy to address the issue of a lack of data is to leverage knowledge from another domain which is known as transfer learning. Instead of training a blank network, using a transfer learning approach can make use of already optimized weights to detect conventional features in the lower layers of the network and only learn the weights in the upper layers. One of the greatest advantages of this technique is that models can train much faster, require less computational power, and converge with sufficiently fewer training data.

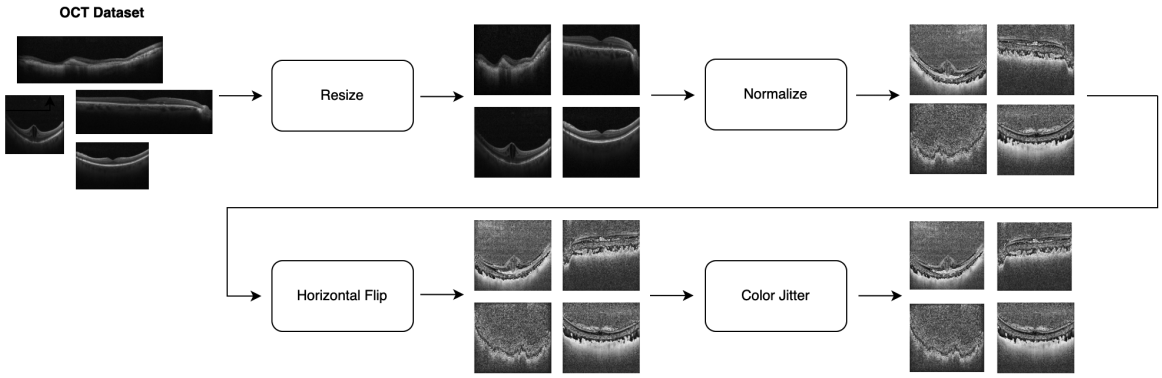

Figure 2. Data preprocessing pipeline.

There are two main types of transfer learning settings: fixed feature extractor and finetuning. A convolutional neural network as a fixed feature extractor is the case of using a pretrained model with all layers' weights frozen and the network is used as a feature extractor. An exception is the last fully connected layer that is required to learn the weights because the pretrained model outputs 1000 classes as per ImageNet and it has to be altered to the number of categories for the given task. The second technique is to finetune the convolutional neural network by learning all the weights in the network by continuing the backpropagation. The finetuning could happen on different levels such as retraining all layers or just some of the higher layers. The intuition behind keeping the earlier layers frozen and only learning the last layers comes from the fact that earlier layers learn more generic features such as edges, corners, colors whereas higher layers focus on more specific features related to the data task. Depending on the size of the dataset and the similarity of the data domain to the ImageNet, different levels of finetuning can be assessed. The smaller the data size is, the more frozen layers should be kept due to overfitting possibilities.

In general, convolutional neural networks provide three scaling dimensions: depth, width, and resolution. The depth is specified by the number of layers of the network, the width is defined by the maximal number of nodes in a layer and the resolution of a CNN is the image resolution passed to the network. Generally, the scaling of neural networks leads to better performance; however, it comes at a cost and brings various complications. For instance, scaling in-depth raises the problem of vanishing gradients, width scaling often causes the accuracy to saturate quicker while resolution scaling comes at a price of computational efficiency and the accuracy improvement decreases in bigger models. Thus it is of high importance to find the optimal scaling along the different dimensions and identify the trade-offs.

In this work, the EfficientNet model architecture is used. The EfficientNet proposed by Tan et al.[7] is established as the current state-of-the-art architecture and addresses the problem of finding the optimal scaling factor for all dimensions of a neural network: depth, width, and resolution. They are the first to express in an empirical manner the relationship among the three dimensions. It has been proven that the balance among the three dimensions is vital for acquiring improved accuracy and efficiency. The EfficientNet architecture is based on scaling up in multiple dimensions rather than in a single one. EfficientNet models consist of a base model, EfficientNet-B0, which is then scaled along the three dimensions and the different scaling factors produce EfficientNets B1 to B7.

In this study, the EfficientNet model architecture is used and particularly an EfficientNet-B4 model is used that is pretrained on the ImageNet dataset. To apply an effective transfer learning strategy, a selection process to identify the number of layers that are required to be retrained for optimal performance is completed. The model achieved the most competitive results when weights from the last four blocks of layers are retrained while the convolutional layers are kept frozen. AdaBound optimizer[10] has been used with a starting learning rate of 0.001 and a final rate of 0.1. AdaBound is an adaptive gradient optimizer from the same family as Adam and RMSprop with a dynamic bound of learning rate which transforms into SGD at the end. Weighted cross-entropy loss is applied to reflect on the irregularly spread class samples. As a regularization technique, early stopping is incorporated to detect the optimal number of epochs to train the model and avoid overfitting.

Semisupervised Learning-Based Solution Details

One of the biggest challenges to develop machine learning solutions in the medical domain is the lack of labeled data samples. Annotating medical data requires medical expertise and is regarded as a major burden because it is a very expensive and time-consuming process. Semisupervised learning is a method that addresses this issue by taking advantage of the abundant amount of unlabeled data along with the labeled samples to enable the building of performant models while requiring less expertise effort and time. Generally, in semisupervised learning, image representations are learned from unlabeled data samples, and then all representations are matched to a label based on the proportion of samples that are labeled.

The primary motivation for this work is the proposed simple framework for contrastive learning of visual representations by GoogleAI, called SimCLR.[8] The SimCLR not only improves significantly previous state-of-the-art self-supervised and semisupervised learning methods and scores 85.8% top-5 accuracy using only 1% of labeled data on the ImageNet dataset but also beats some of the supervised learning methods.

Contrastive learning is learning based on distinctiveness and is at the core of the SimCLR framework. It attempts to find what makes two pairs of images similar or dissimilar. SimCLR learns the visual representations of the images by taking variously augmented versions of the same data sample and maximizing the agreement between the augmented samples by applying a contrastive loss.

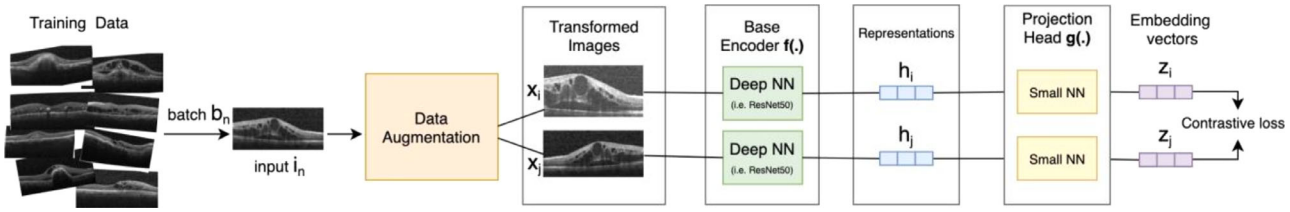

Figure 3. SimCLR framework workflow.

The SimCLR framework flow is presented in Figure 3. The first step is to generate batches of size N from the training samples where $N \in [128, 8192]$. An augmentation function is defined to get an input of an image and apply a random combination of image transformations. For each data sample in the batch, a random data augmentation function is applied to produce a pair of two images x_i and x_j for each input resulting in total $2N$ total samples. Then, each of the augmented images is passed into an encoder to extract the image representation vectors h_i and h_j. The SimCLR framework is not restricted and allows any choice of neural architecture as a base encoder. Then, the h_i and h_j image vectors are passed into a neural network called a projection head to apply non-linear transformation and map them into z_i and z_j representations. This produces an embedding vector z for each of the augmented images in the batch. Next, the similarity between each two augmented images is calculated using cosine similarity. This score shows how similar or dissimilar each image pairs are. The assumption is that the similarity between augmented images of class A will be high while the similarity between samples of class A and class B will be lower. Finally, a contrastive loss is calculated for every two augmented pairs. The task is to maximize the similarity between the z_i and z_j representations of the same image. To evaluate the learned representation z_i and z_j, a commonly used linear evaluation protocol is followed where the base network is frozen and a linear classifier is trained on top of it.

The SimCLR framework is used in this work. Following the protocol of SimCLR, the data augmentations applied in all of the experiments are random cropping, resizing, random flipping, color distortion, and Gaussian blur. The ResNet-50 architecture is used as a base encoder and an MLP with one hidden layer as a projection head. The performance of the SimCLR has been assessed in various configuration settings to find the optimal hyperparameters. The base encoder is trained for 1000 epochs with a batch size of 256. Training fewer epochs does not lead to convergence, whereas training more epochs does not further benefit the results. Although using a bigger batch size is almost certainly expected to improve performance, in this work 256 is selected due to computational limitations. The NT-Xent is used for a loss function. because the selected batch size in this study is comparatively small (256), Adam is used as an optimizer in contrast to the SimCLR paper where LARS is adopted. LARS is advantageous to stabilize training with bigger batch sizes which is not beneficial in this work. After the training of the base encoder, the labels of a random number of data samples are revealed (either 1% or 10% of data samples) and then a linear classifier is trained on top of the learned representations from the base encoder.

Results

Supervised Transfer Learning Experimental Results

Table 2 shows the comparative quantitative results for the experiments performed on different dataset variations using the EfficientNet-B4 model architecture.

The proposed EfficientNet-B4 model trained on the original dataset achieved 0.9812 accuracy with a sensitivity score of 0.981 and a specificity score of 0.9936. The performance of the EfficientNet-B4 trained on the limited dataset is similar to the network trained

Table 2. Comparative Quantitative Results of the EfficientNet-B4 Network in Different Dataset Variations

Metric	EN-b4 Original	EN-b4 Limited	EN-b4 Mini
Dataset	original	limited	mini
Accuracy	0.9812	0.9762	0.8363
Loss	0.0558	0.0907	0.7011
Sensitivity (TPR)	0.981	0.973	0.683
Specificity (TNR)	0.9936	0.991	0.8943
F1 Score	0.9806	0.9729	0.8276
CKS	0.975	0.964	0.7157
MCC	0.9742	0.9649	0.6890

TPR, true positive rate; TNR, true negative rate; CKS, Cohen kappa score; MCC, Matthews correlation coefficient.

on the original dataset. The EfficientNet-B4 on the limited data obtained an accuracy of 0.9762 with 0.973 and 0.991 rates for sensitivity and specificity. A high sensitivity rate demonstrates high abilities of the model to correctly identify scans with eye pathologies (true positive rate), whereas the specificity rate shows that the model learned to recognize healthy eyes (true negative rates). A low false negative rate is extremely critical for medical diagnostics systems. False negative results are more detrimental than false positive because the false negative result implies that the model failed to refer the patient accordingly which could lead to the patient not being treated and ultimately causing fatalities. Therefore a high sensitivity rate is of huge importance in medical diagnosis applications. The model trained on the mini dataset scored 0.8363 accuracy, 0.683 sensitivity, and 0.8943 specificity.

Since the accuracy score assumes that both false negatives and false positives have similar costs, which is not the case in the medical diagnosis; the F1 Score is a metric that considers both false negatives and false positives. The classifiers trained on the original and limited dataset obtained F1 Scores of 0.9806 and 0.9729, respectively, whereas the model on the mini set achieved a rate of 0.8276.

The Cohen kappa score is an evaluation metric that considers imbalanced data and shows how much better the model performed compared to a random classifier given the frequencies of each class.[11] The networks developed on the original and the limited dataset achieved a kappa score of 0.975 and 0.964, respectively, and the classifier on the mini dataset acquired a score of 0.7157.

Similar results to the Cohen kappa score are observed when comparing the models' performance based on the Matthews correlation coefficient score with the model trained on the original data achieving 0.9747, the limited one scoring 0.9649 and the mini with 0.6890.

Tables 3 and 4 present the distribution of the evaluation metrics for each of the four OCT categories—CNV, DME, drusen, and normal.

Table 3. Summary of the Evaluation Metrics Per OCT Category for the EfficientNet-B4 Developed on the Original Dataset

Category	Accuracy	Precision	Recall	F1-Score
CNV	0.988	0.9643	0.992	0.974
DME	0.996	0.986	0.9935	0.9854
Drusen	0.96	0.982	0.9642	0.972
Normal	0.98	0.996	0.986	0.994

Semisupervised Learning Experimental Results

Table 5 presents the conducted experiments using the SimCLR framework. Both the limited dataset and the mini dataset with 1% and 10% labeled data have been used for the experiments. One of the goals was to examine the impact of the number of total samples and the percentage of labeled samples on the generalization abilities of the network.

The classifier trained on the limited dataset with 10% labeled data achieved an accuracy of 0.946. As discussed earlier, a low false negative rate is extremely important in the medical domain. The classifier achieved a sensitivity score of 0.941 and a specificity score of 0.9825. When training a semisupervised model on the limited dataset with 1% labeled data, the classifier achieved a 0.8117 accuracy rate. The model scored a rate of 0.8063 for sensitivity and 0.9120 for specificity. Another experiment on the mini dataset with 10% labeled data has been done to examine the model's abilities to identify eye disorders when much less data are available. The model achieved an accuracy of 0.8435.

Table 6 shows the performance of the SimCLR network developed on the 10% labeled data samples across the four OCT dataset categories—CNV, DME, drusen, and normal.

Table 4. Summary of the Evaluation Metrics Per OCT Category for the EfficientNet-B4 Developed on the Limited Dataset

Category	Accuracy	Precision	Recall	F1-Score
CNV	0.998	0.9504	0.9960	0.9727
DME	0.992	0.9880	0.9920	0.9900
Drusen	0.952	0.9754	0.9520	0.9636
Normal	0.968	0.9959	0.9680	0.9817

Table 5. Comparative Quantitative Results of the SimCLR Performance in Different Dataset Variations

Metric	SimCLR Limited 10%	SimCLR Limited 1%	SimCLR Mini 10%
Dataset	limited	limited	mini
Labeled data	10%	1%	10%
Accuracy	0.946	0.8117	0.8435
Loss	0.2142	0.6271	0.6829
Sensitivity (TPR)	0.941	0.8063	0.8127
Specificity (TNR)	0.9825	0.9120	0.9223

Table 6. Summary of the Evaluation Metrics Per OCT Category for the SimCLR Developed on the 10% Labeled Limited Dataset

Category	Accuracy	Precision	Recall	F1-Score
CNV	0.976	0.8971	0.9760	0.9349
DME	0.996	0.9542	1.0000	0.9766
Drusen	0.876	0.9520	0.8720	0.9102
Normal	0.936	0.9873	0.9360	0.9610

Discussion

Transfer learning applications in image classification have been explored recently in the OCT domain in a range of publications.[12–15] Kermany et al.[12] explored the application of transfer learning with ImageNet weights and Inception V3 model architecture. As part of their work they also published their dataset, which is currently the largest publicly available dataset and is the one used in this work. The proposed classifier by Kermany et al.[12] acquired an accuracy of 0.966, sensitivity of 0.978 and specificity of 0.974. Another example work using a transfer learning-based approach is the Li et al.[15] proposed solution to detect OCT pathologies using a pretrained VGG-16 with ImageNet weights. Accuracy of 0.986 was achieved with 0.978 sensitivity and 0.994 specificity.

An observation over the reviewed materials is that although in most cases the datasets are highly skewed, there is no evidence that it has been taken under consideration during the model development or training phase. Therefore the data imbalance problem is addressed thoughtfully in this work. Another common issue that has not been discussed by most of the studies is the importance of splitting the data correctly. OCT scans present 3D volumes of images and different dataset partitions might contain multiple scans of the same patient. To clarify, the OCT scans belonging to the same patient are often quite identical and their placement in different data partitions such as train, test, validation, will introduce bias to the model, lead to overfitting and produce misleading performance results. Thus, in this work, it is carefully considered to split the data precisely as it affects both supervised based learning and semisupervised learning.

In previous works, semisupervised learning approaches for image classification have been investigated in medical domains such as brain tumor,[4] breast cancer,[5] and chest radiography,[6] but no studies researched the semisupervised learning strategy in the OCT data domain. In all these studies, the development of novel semisupervised frameworks to classify medical images has been explored. The proposed SimCLR[8] framework improved significantly previous state-of-the-art semisupervised learning methods and also defeated some of the supervised learning methods. It scores 85.8% top-5 accuracy using only 1% of labeled data on the ImageNet dataset.

Table 7 describes the main experiments completed in this study both in fully supervised and semisupervised settings on the three different versions of the OCT dataset—original, limited, and mini. The table also outlines the impact of the proportion of provided labeled samples on the downstream performance.

Both V2 and V7 experiments are conducted on the limited OCT dataset with 100% labeled data. In the V2 experiment, the transfer learning approach is taken whereas in V7 semisupervised approach is applied. As can be seen, the performance of the models in both experiments is nearly equivalent. The fully supervised model achieved an accuracy of 0.9762, whereas the semisupervised-based network scored 0.9788 accuracy. The other evaluation metrics follow the same trend as the accuracy and are very similar. This demonstrates that the results from the semisupervised approach (V7) are comparable to the fully supervised one (V2), and

Table 7. Comparative Quantitative Results of the Performance of the EfficientNet-B4 Network and the SimCLR Framework in Different Dataset Variations

Metric	EN-b4 Original	EN-b4 Limited	EN-b4 Mini	SimCLR Limited 10%	SimCLR Limited 1%	SimCLR Mini 10%	SimCLR Limited 100%
Name	V1	V2	V3	V4	V5	V6	V7
Dataset	original	limited	mini	limited	limited	mini	limited
Label data	100%	100%	100%	10%	1%	10%	100%
Accuracy	0.9812	0.9762	0.8363	0.946	0.8117	0.8435	0.9788
Loss	0.0558	0.0907	0.7011	0.2142	0.6271	0.6829	0.2382
Sensitivity	0.981	0.973	0.683	0.941	0.8063	0.8127	0.9788
Specificity	0.9936	0.991	0.8943	0.9825	0.9120	0.9223	0.9937

both frameworks learned the distinguishable features of the data domain and developed an ability to discriminates different eye pathologies.

The results achieved by the semisupervised classifier trained with 10% labels (V4) are comparable to the results achieved by the fully supervised classifier trained on 100% labeled data using transfer learning (V2). Both experiments are conducted on the same limited dataset but on different proportions of labeled data. Using transfer learning and training on a fully labeled dataset yielded an accuracy of 0.9762, whereas training with just 10% labeled data and using semisupervised learning acquired a very closed accuracy score of 0.946. The same performance pattern is observed when comparing the other evaluation metrics such as sensitivity, specificity, F1 score, Cohen kappa score and Matthews correlation coefficient. Although the semisupervised model is provided with much fewer data samples to learn from, it demonstrated extremely high abilities to identify the different OCT categories. This demonstrates the power of semisupervised learning and its tremendous potential to be used in image classification.

Another interesting comparison to look into is the experiments V3 and V6. Both are executed on the mini dataset. Although in the V3 experiment a supervised transfer learning approach is taken, in the V6 experiment the semisupervised learning is applied with merely 10% labeled data. The accuracy results achieved by the two experiments are very close, 0.8363 and 0.8435 for the supervised and semisupervised approaches, respectively. It can be also seen that not only the accuracy but also the other evaluation performance metrics are alike. This suggests that both models, V3 and V6, developed identical learning capabilities even though the semisupervised-based network is provided 90% less labeled data. The same performance trend is observed by V2 and V4 experiments as discussed earlier. The obtained results illustrate that the semisupervised models have a huge potential to be capable of extracting information and learning distinguishable features from an insufficient amount of labeled inputs.

The experiments done on the limited dataset with 1% labeled data (V5) and the mini dataset with 10% (V6) accomplished similar results. In both experiments, the models are trained on a very limited number of labeled samples. Although the labeled data represented a different ratio of the total data provided in the two experiments (1% and 10% labeled samples, respectively), the actual number of labeled inputs is similar and so are the results. This leads to the assumption that the number of provided unlabeled samples does not contribute highly to the learning process of the models. The same trend is demonstrated in the experiments on the limited dataset with 1% and 10% labeled samples where the model trained on more labeled images acquired better generalization abilities. The experiment on the mini dataset also shows that the amount of labeled data plays a vital role in the learning process of a neural network.

In summary, the semisupervised learning-based experiments demonstrated that the semisupervised networks are equally performant to fully-supervised ones. Although, semisupervised models are provided with substantially fewer labeled samples, they are capable of learning distinct data features and develop competitive generalization abilities. Some of the limitations of this study are related to the diversity of the data. The models are trained on a dataset consists of images acquired via the Heidelberg Spectralis (Heidelberg Engineering, Heidelberg, Germany) imaging platform[12]; however, other OCT scanner manufacturers also exist (e.g., Zeiss, Oberkochen, Germany, and Optovue, Fremont, CA, USA). Although different producers have made the resulted OCT scans reasonably consistent, it will be beneficial to collect a dataset comprising of scans of various imaging platform manufacturers to validate the abilities of the neural networks and further improve their generalization abilities.

Another limitation of the work is that it has not addressed the availability of OCT volumes. In essence, the OCT scans represent 3D volumes. It will be interesting to explore which of the slices in the volume gives the most substantial information to the model regarding the category and if there are particular scans that the algorithms are more interested in (e.g. the frontest scan of the volume, the most back scan, or some in the middle of the volume). For this idea to be further explored, a dataset containing patients' volume scans needs to be collected.

In the medical domain, data are often imbalanced with the minority classes in most cases being the ones of interest. Because the undersampling technique has been already explored as part of this work when constructing the limited dataset, it will be interesting to apply the oversampling strategy. The application of Generative Adversarial Networks to generate synthetic data samples is another limitation of this study that it will be beneficial to explore. The effect of the synthetic data on the classification abilities of the models can be analyzed and compared to the results that were obtained without the addition of the synthetic data.

Interpretability

Another big challenge in the deep learning field is to provide the interpretability of the models. Machine learning algorithms are often treated as black boxes

because of the limited insight behind their decision making. This leads to a low acceptance rate by the general public and criticism regarding their deployment in production and use in important fields. Furthermore, the adoption of machine learning systems in critical fields such as medical diagnosis requires well defined and precise measures which state how the systems will be held accountable and interpretable. This section of the study aims to explore methods of explaining and interpreting the developed models.

Local interpretable model-agnostic explanation (LIME)[16] is a model-agnostic technique that implements a local surrogates model. Local surrogate models are interpretable models such as linear regression and decision trees. The aim of the surrogate models is to approximate the underlying model's prediction. LIME focuses on explaining individual predictions by training local surrogate models rather than global ones. LIME helps to understand which region of the image contributed the most to the model classification decision and how different superpixels (set of pixels) of the image affect the final predictions. LIME creates explanations for individual samples from the dataset. First, LIME constructs a dataset from randomly permuted samples of a particular image by turning on and off superpixels. Then, each of the permuted images is assigned a classification label by using the trained machine learning model. Next, the samples in the newly generated dataset are weighted based on the data sample that is being explained. The final step is to fit a local surrogate model to explain the predictions of the classifier.

LIME is applied to explain the predictions of two of the main models developed as part of this study, namely the supervised EfficientNet-B4 model and the semisupervised SimCLR trained on 10% labeled data corresponding to experiment versions V1 and V4 as per Table 7. Both models are developed on the same limited dataset as explained in detail in the Discussion section.

Another explainability technique applied in this work is the class activation maps (CAM). Class activation maps (CAM) are a simple yet effective technique to retrieve the image regions which a neural network uses to detect a particular class.[17] To generate the class activation maps, the global average pooling layer in the classifier is used. CAM depicts the region in the image used by CNN to assign a certain label to an image. Before the final layer of the neural network, which is often a SoftMax layer in the case of classification, a global average pooling is performed on the convolutional feature maps and these features are fed into the fully connected layer. This way the important regions of the image can be detected by projecting back the weights of the output layer to the convolutional feature maps acquired by the last convolutional layer. In other words, CAM reveals the category-specific discriminative region. CAM is particularly beneficial to interpret the prediction decision made by the neural network and gaining insight into the decision process of the neural network.

Drusen is an eye disorder belonging to AMD same as CNV. Drusen represents tiny irregularities built between the retinal pigment epithelium and Bruch membrane layers of the retina. In Figure 4 are depicted scans of class drusen, the corresponding confidence scores of the models for the top 1 feature explanation from LIME and the class activation maps. Both V1 and

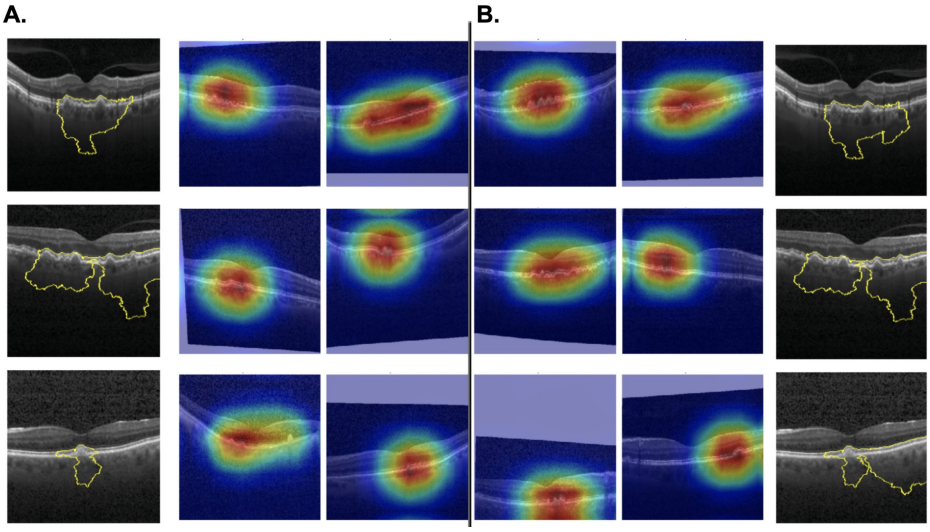

Figure 4. LIME and CAM visualizations for drusen category for (A) EfficientNet-B4 (V1) and (B) SimCLR (V4).

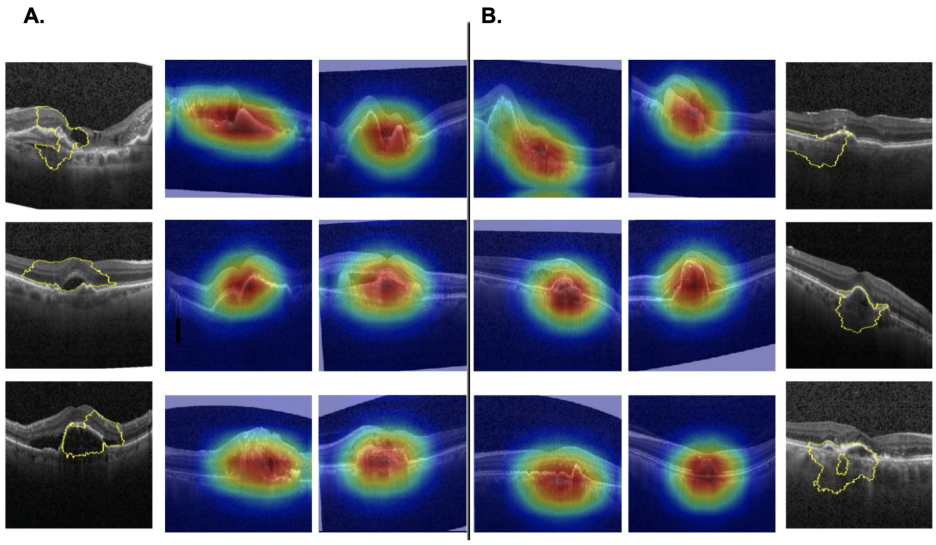

Figure 5. LIME and CAM visualizations for CNV category for (A) EfficientNet-B4 (V1) and (B) SimCLR (V4).

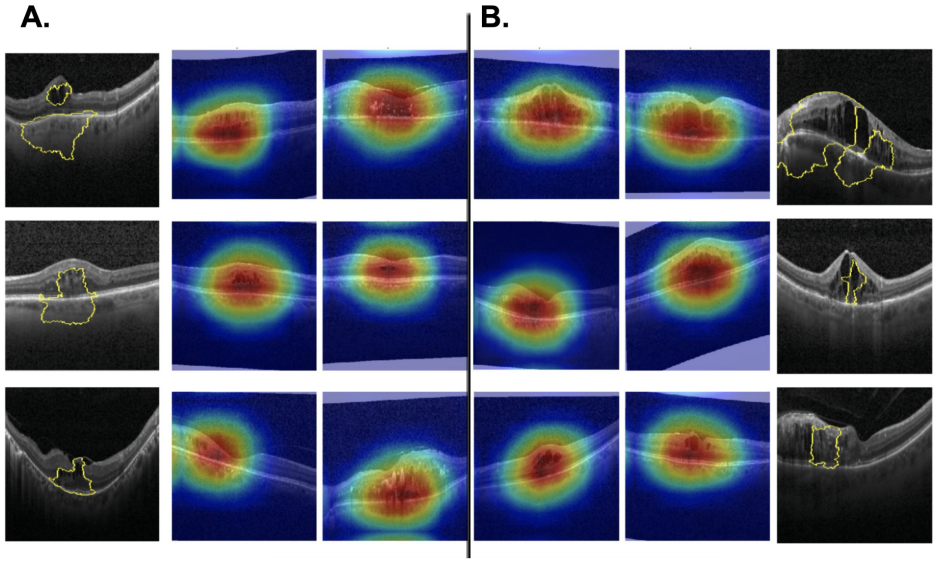

Figure 6. LIME and CAM Visualizations for DME category for (A) EfficientNet-B4 (V1) and (B) SimCLR (V4).

V4 models focus on the extracellular material accumulated between retinal layers, which indicates that the models learned correctly how to discriminate drusen disorder. Also, as can be seen, the two models show practically the same images' segments, which is proof that the semisupervised model has accurately learned the features of the category.

When abnormal blood vessels grow inside the retina and begin leaking fluid, then age-related macular degeneration (AMD) has progressed into wet AMD, which is also referred to as choroidal neovascularization or CNV. Figure 5 illustrates the superpixels that LIME found to be of the greatest importance for the models for classification prediction and the regions that CAM detected to have contributed to the classification label the most. The LIME segments match precisely the visibly grown blood vessels associated with CNV. This demonstrates that the neural networks have learned the feature characteristics of the CNV category.

DME is another eye disorder associated with patients with underlying diabetes. It is associated with fluid accumulation in the macula and retinal thickening. DME can be recognized on a scan by an accumulated subretinal fluid. Figure 6 presents the corresponding LIME segments and CAM visualizations that are clearly around the regions of the images that represent accumulated fluid between the layers. This is a confirmation that the network developed the ability to recognize DME and to discriminate its prime features.

Conclusion

Medical domains are generally associated with limited data samples. It is particularly difficult to acquire a sufficient amount of verified and labeled data in the medical field because it requires domain-specific expertise. Professional knowledge comes at high expenses and is a very lengthy process. In numerous cases, datasets consist of more unlabeled samples than labeled. Limited labeled data are often one of the main obstacles to developing machine learning solutions in the health sector. Therefore the promises of semisupervised learning in the field are huge. In this work, we proposed two machine learning solutions—one based on supervised transfer learning and another one based on semisupervised learning. We compared the developed supervised transfer learning models to the semisupervised classifiers to assess the performance. This helped to validate the potential of using semisupervised learning in domains with scarce amounts of labeled data. The experiments demonstrated that the semisupervised machine learning solution for medical diagnosis is capable to gain in-depth knowledge through limited labeled data and minimum supervision. The applied interpretability techniques further explained the effectiveness of the proposed solutions.

Acknowledgments

The authors thank the Science and Technology Unit, King Abdulaziz University for technical support.

Supported by the National Plan for Science, Technology and Innovation (MAARIFAH), King Abdulaziz City for Science and Technology, the Kingdom of Saudi Arabia, award number (10-INF1262- 03).

Disclosure: **L. Al Turk,** None; **D. Georgieva,** None; **H. Alsawadi,** None; **S. Wang,** None; **P. Krause,** None; **H. Alsawadi,** None; **A.Z. Alshamrani,** None; **G.M. Saleh,** None; **H.L. Tang,** None

* LAT and DG are joint first authors.

References

1. Swanson DHE, Huang D. Ophthalmic OCT reaches $1 billion per year. *Retin Physician*. 2011;8(4):45:5859.
2. Bourne RRA, Flaxman SR, Braithwaite T, et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *Lancet Glob Health*. 2017;5:e888–e897.
3. Cheplygina V, de Bruijne M, Pluim JPW. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med Image Anal*. 2019;54:280–296.
4. Ge C, Gu IY-H, Jakola AS, Yang J. Deep semi-supervised learning for brain tumor classification. *BMC Med Imaging*. 2020;20:87.
5. Sun W, Tseng T-LB, Zhang J, Qian W. Computerized breast cancer analysis system using three stage semi-supervised learning method. *Comput Methods Programs Biomed*. 2016;135:77–88.
6. Aviles-Rivero AI, Papadakis N, Li R, Sellars P, Fan Q, Tan RT, Schönlieb C-B. GraphxNET chest x-ray classification under extreme minimal supervision. arXiv preprint arXiv:1907.10085.
7. Tan M, Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. Proceedings of the 36th International Conference on Machine Learning. *PMLR*. 2019;97:6105–6114.
8. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. Proceedings of the 36th International Conference on Machine Learning. *PMLR*. 2020;119:1597–1607.
9. Kermany D, Zhang K, Goldbaum M. Large dataset of labeled optical coherence tomography (OCT) and chest x-ray images. Mendeley Data [serial online], https://doi.org/10.17632/rscbjbr9sj.3.
10. Luo L, Xiong Y, Liu Y, Sun X. Adaptive gradient methods with dynamic bound of learning rate. arXiv preprint arXiv:1902.09843.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159.
12. Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172:1122–1131.e9.
13. Lu W, Tong Y, Yu Y, Xing Y, Chen C, Shen Y. Deep learning-based automated classification of multi-categorical abnormalities from optical coherence tomography images. *Transl Vis Sci Technol*. 2018;7:41.
14. Karri SPK, Chakraborty D, Chatterjee J. Transfer learning based classification of optical coherence tomography images with diabetic macular edema

and dry age-related macular degeneration. *Biomed Opt Express*. 2017;8:579.

15. Li F, Chen H, Liu Z, Zhang X, Wu Z. Fully automated detection of retinal disorders by image-based deep learning. *Graefes Arch Clin Exp Ophthalmol*. 2019;257:495–505.
16. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016:1135–1144.
17. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:2921–2929.