



OPEN Evaluating deep learning models for classifying OCT images with limited data and noisy labels

Aleksandar Miladinović^{1✉}, Alessandro Biscontin¹, Miloš Ajčević², Simone Krešević², Agostino Accardo², Dario Marangoni³, Daniele Tognetto³ & Leandro Infrerera³

The use of deep learning for OCT image classification could enhance the diagnosis and monitoring of retinal diseases. However, challenges like variability in retinal abnormalities, noise, and artifacts in OCT images limit its clinical use. Our study aimed to evaluate the performance of various deep learning (DL) architectures in classifying retinal pathologies versus healthy cases based on OCT images, under data scarcity and label noise. We examined five DL architectures: ResNet18, ResNet34, ResNet50, VGG16, and InceptionV3. Fine-tuning of the pre-trained models was conducted on 5526 OCT images and reduced subsets down to 21 images to evaluate performance under data scarcity. The performance of models fine-tuned on subsets with label noise levels of 10%, 15%, and 20% was evaluated. All DL architectures achieved high classification accuracy (> 90%) with training sets of 345 or more images. InceptionV3 achieved the highest classification accuracy (99%) when trained on the entire training set. However, classification accuracy decreased and variability increased as sample size decreased. Label noise significantly affected model accuracy. Compensating for labeling errors of 10%, 15%, and 20% requires approximately 4, 9, and 14 times more images in the training set to reach the performance of 345 correctly labeled images. The results showed that DL models fine-tuned on sets of 345 or more OCT images can accurately classify retinal pathologies versus healthy controls. Our findings highlight that while mislabeling errors significantly impact classification performance in OCT analysis, this can be effectively mitigated by increasing the training sample size. By addressing data scarcity and labeling errors, our research aims to improve the real-world application and accuracy of retinal disease management.

Keywords OCT, CNN, Deep learning, Retina, Retinal imaging

Despite progress in medical imaging technologies, Optical Coherence Tomography (OCT) stands out as the premier tool for evaluating retinal conditions in the eye^{1,2}. This non-invasive imaging technique is essential for diagnosing retinopathy, offering detailed cross-sectional views and high-resolution images of the retina's various layers^{3,4}. OCT significantly aids in therapeutic decisions and the diagnosis of multiple retinal disorders, enabling ophthalmologists to discern the retina's layers with precision⁵. This imaging modality allows for detailed visualization of the retinal architecture, enabling early detection and monitoring of various ocular conditions such as age-related macular degeneration^{6–9} (AMD), diabetic retinopathy^{10–12}, macular hole¹³ (MH) and glaucoma^{14,15}.

The non-invasive nature of OCT, coupled with its ability to produce real-time images, has made it a standard tool in modern ophthalmic practice¹⁶.

This advanced imaging technology plays a key role in identifying both the normal structure of the retina and signs of retinal pathologies, allowing for timely diagnosis and intervention in several retinal conditions¹⁷.

One of the retinal signs detectable via OCT is intraretinal fluid (IF), which appears as hyporeflective areas within the neurosensory retina, as well as subretinal fluid (SF), a clear or lipid-rich exudate that accumulates between the neurosensory retina and the retinal pigment epithelium¹⁸. These fluid accumulations are often linked to conditions like diabetic retinopathy, exudative AMD or retinal vein occlusion. Additionally, drusen (D) - subretinal deposits found beneath the retinal pigment epithelium - are closely associated with AMD and serve as early indicators of disease progression¹⁹. In cases of exudative AMD, a common complication is macular

¹Institute for Maternal and Child Health IRCCS "Burlo Garofolo", Via dell'Istria 1, Trieste 34100, Italy.

²Department of Engineering and Architecture, University of Trieste, Trieste, Italy. ³Department of Medicine, Surgery and Health Sciences, Ophthalmology Clinic, University of Trieste, Trieste, Italy. ✉email: aleksandar.miladinovic@burlo.trieste.it

neovascularization (MNV), which is triggered by the overproduction of vascular endothelial growth factor in response to localized ischemia and inflammation²⁰.

In exudative AMD, macular neovascularization (MNV) commonly occurs as a result of increased vascular endothelial growth factor production in response to localized ischemia and inflammation²⁰. This can lead to disruptions in the retinal pigment epithelium (RPE), and while the OCT signal normally penetrates the RPE minimally in a healthy retina, increased choroidal backscatter (BS) indicates a compromise of the RPE's usual barrier to the OCT laser²¹. Another age-related condition is vitreomacular adhesion (VMA), which happens when the vitreous starts to detach from the retina but remains partially attached at certain points near the fovea²². Besides these conditions, other retinal signs include epiretinal membranes (ERM), also known as macular puckers or cellophane maculopathy. This is a semi-translucent fibrous tissue that forms on the inner surface of the retina, particularly affecting the macula and often leading to vision problems²³. Another key pathological sign is the presence of macular holes (MH), characterized by full or partial defects in the central macular layers, which can significantly impair vision²⁴. Representative OCT images illustrating a normal retina and various retinal abnormalities are shown in Fig. 1.

To better identify and analyze such conditions, the advent of deep learning (DL), particularly pre-trained Convolutional Neural Networks (CNNs), has further improved OCT image analysis²⁵. DL models have demonstrated remarkable capabilities in various image-processing tasks, including classification, segmentation, and anomaly detection^{26,27}. These models are particularly adept at identifying intricate patterns and features within images, which are often challenging for traditional image processing techniques²⁶.

In the context of OCT, DL offers several advantages. First, it provides automated, efficient, and accurate solutions for detecting retinal abnormalities^{28,29}. By leveraging large datasets of OCT images, DL models can learn to recognize subtle variations and anomalies that may indicate disease^{28,30}. This capability is crucial for early diagnosis and treatment, potentially preventing vision loss in patients. Second, DL models can process large volumes of data rapidly, providing consistent and objective analyses that enhance clinical decision-making^{31,32}. Moreover, DL-based imaging is increasingly used, to predict outcomes and choose optimal therapeutic strategies in personalized medicine³³ and improve overall diagnosis in clinical settings where time and resources are



Fig. 1. Representative OCT images illustrating a normal retina and various retinal abnormalities, including intraretinal fluid (IF), subretinal fluid (SF), drusen (D), macular neovascularization (MNV), vitreomacular adhesion (VMA), epiretinal membrane (ERM), macular holes (MH), and choroidal backscattering (BS).

limited³⁴. Despite numerous publications highlighting their advantages, the application of DL in OCT analysis is not without challenges, and these models are not yet widely applied in clinical settings³⁵.

Considering the complexity and variability of retina abnormalities, the subtle changes of retinal layers in retinopathies, and pitfalls of OCT images which can suffer from noise artifacts³⁶ and variability in acquisition settings³⁵, a large volume of labeled data is needed to train DL models effectively. In addition, accurate data labeling is essential for achieving high classification performance and generalizability²⁶. However, the process of labeling OCT images is labor-intensive and requires expertise. Human error in labeling introduces inconsistencies and biases, potentially impacting the performance of trained models²⁸. Studies indicate that human error in labeling medical images can be up to 17%³⁷, depending on factors such as image complexity, interobserver variability, and fatigue³⁸. Such errors can propagate throughout the training process, affecting the performance of AI-based support systems³⁹.

Our study aimed to systematically evaluate the performance of various deep learning (DL) architectures for classifying retinal pathologies versus healthy cases based on OCT images, under conditions of data scarcity and label noise. Specifically, we aimed to identify the minimal data requirements for effective model training, investigate the impact of label noise on model performance, and determine how different levels of labeling errors affect the number of images needed to achieve clinically relevant accuracy. Our ultimate goal was to provide insights that can inform the creation of clinical guidelines related to the detection, diagnosis, and management of retinal diseases using DL models, thereby enhancing their reliability and effectiveness in real-world clinical settings.

Materials and methods

Study design

This retrospective observational study was conducted at the University Eye Clinic of Trieste, following the ethical guidelines of the Declaration of Helsinki. The study protocol received approval from the Regional Ethics Committee (CEUR) of Friuli Venezia Giulia, Italy (protocol number 17094/2022). All participants gave their informed consent for the use of their data.

Data collection

We retrospectively evaluated anonymized OCT scans obtained using an A-line scan protocol with a length of 9.0 mm. The images were acquired with the Spectralis OCT system (Heidelberg Engineering, Heidelberg, Germany), with a central wavelength of 815 nm, an axial resolution of 3.9 $\mu\text{m}/\text{pixel}$, a lateral resolution of 5.7 $\mu\text{m}/\text{pixel}$, and an image dimension of 768 $\times$ 496 pixels. The study included both horizontal and vertical line scans centered on the fovea. These scans were collected from individuals aged 18 to 95 years between January 2017 and September 2022.

Two experienced retinal specialists (LI and DM) thoroughly examined and labeled all the images. Images of poor quality, those not centered on the fovea, and scans where the specialists did not agree on the diagnosis were excluded from the dataset.

Cases were divided into two groups: healthy and those with retinal pathologies. The pathological group included scans showing signs of ERM, IF, SF, D, MNV, VMA, MH, and BS. The control group comprised cases with no retinal abnormalities in their OCT scans. The total number of healthy scans was randomly selected to match the number of scans in the pathological group. This equal selection helps avoid model bias toward either normal or pathological cases. Images that did not meet the quality threshold (Spectralis Quality parameter < 23) were excluded from the study. The final dataset included 6240 images (3120 pathological and 3120 healthy).

Comparative analysis of deep learning architectures

We employed a set of well-established DL pre-trained architectures for our study, namely ResNet18, ResNet34, ResNet50, VGG16, and InceptionV3^{40–42}.

Initially, we loaded pre-trained models of ResNet18, ResNet34, ResNet50, VGG16, and InceptionV3, all of which had been trained on the ImageNet dataset⁴³. Subsequently, training, validation, and test set images were defined, containing 5526, 614, and 100 images, respectively. Each pre-trained model was fine-tuned on the entire training set, as well as, on reduced subsets, in order to evaluate the performance of the architecture under data scarcity. These subsets were constructed by dividing the entire training set into progressively smaller sizes, following a geometric progression, with the formula:

$$\text{subset size} = \frac{\text{entire dataset}}{2^n}$$

where n ranged from 1 to 8, resulting in train subset sizes of 5526, 2763, 1381, 690, 345, 172, 86, 43, and 21 images.

Considering the two-class classification problem of our study, each of the pre-trained DL architectures was adapted to function as a binary classifier. This modification involved removing the final fully connected layer and replacing it with a single neuron, which facilitated binary classification tasks⁴⁴.

To train the modified networks, we employed the Adaptive Moment Estimation Algorithm (ADAM) with a fixed learning rate of 0.0001⁴⁵. During the training process, we utilized the Binary Cross Entropy with Logits Loss function as the loss criterion. Such a function is designed for binary classification tasks, offering several advantages over traditional loss functions such as Mean Squared Error⁴⁶.

For each DL architecture and training subset, five training runs were performed. In each run, the model with the minimal validation loss was selected as the best model. Subsequently, the five best models for each DL

architecture and training subset combination were evaluated on the test set using classification accuracy. The overall procedure is summarized in Fig. 2.

Evaluation of the impact of noisy labeling

To evaluate the impact of noisy labeling on classification performance we introduced labeling errors of 10%, 15%, and 20% to simulate the possible real-world scenarios. For each labeling error rate, the training procedure, described in the previous subsection, was repeated for all considered architectures and training subsets in which the mean classification accuracy, without mislabeling, was higher than 90%. This evaluation was performed to assess how noise in the labels, which is common in clinical data due to variability in human annotations, affects the performance of OCT DL models. The additional analysis of the obtained test set accuracies was performed to identify the additional number of images needed in the training set to achieve the classification performance of models trained on the non-misabeled training set.

This evaluation was repeated across all architectures, ensuring a comprehensive analysis and comparison of their resilience to labeling errors and dataset size variations. This allowed us to ascertain the robustness of each architecture to label noise and optimize dataset size for reliable model performance^{47–49}.

Results

Impact of training set size on model performance

In Fig. 3. are reported means $\pm$ SD values of the classification accuracy on the test set obtained for each DL architecture and train subset sample size.

All considered architectures showed high classification accuracy (> 90%) for models trained on 345 or more images.

For the larger sample size subsets, no considerable differences between different considered DL architectures were observed. However, the best performance was achieved by InceptionV3 trained on the entire training set (5526 images). In Fig. 4, we report the GradCam of a pathological OCT image as well as the corresponding GradCam obtained on the InceptionV3 model trained on 345 and 5226 images, showing a similar decision pattern.

Overall, a progressive decrease in average classification performance and an increase in variability of classification accuracy were observed as sample size decreased.

In particular, all architectures achieve accuracy greater than 95% when trained on the largest dataset (5526 images), with a slight decline as the training set size decreases, maintaining accuracy above 95% for dataset sizes down to 690 images. The SD of classification accuracy is minimal for larger datasets, indicating consistent performance. A pronounced drop in accuracy is observed as the dataset size is further reduced, with varied performance around 90–95% for 345 images, where ResNet34 and ResNet50 perform slightly better.

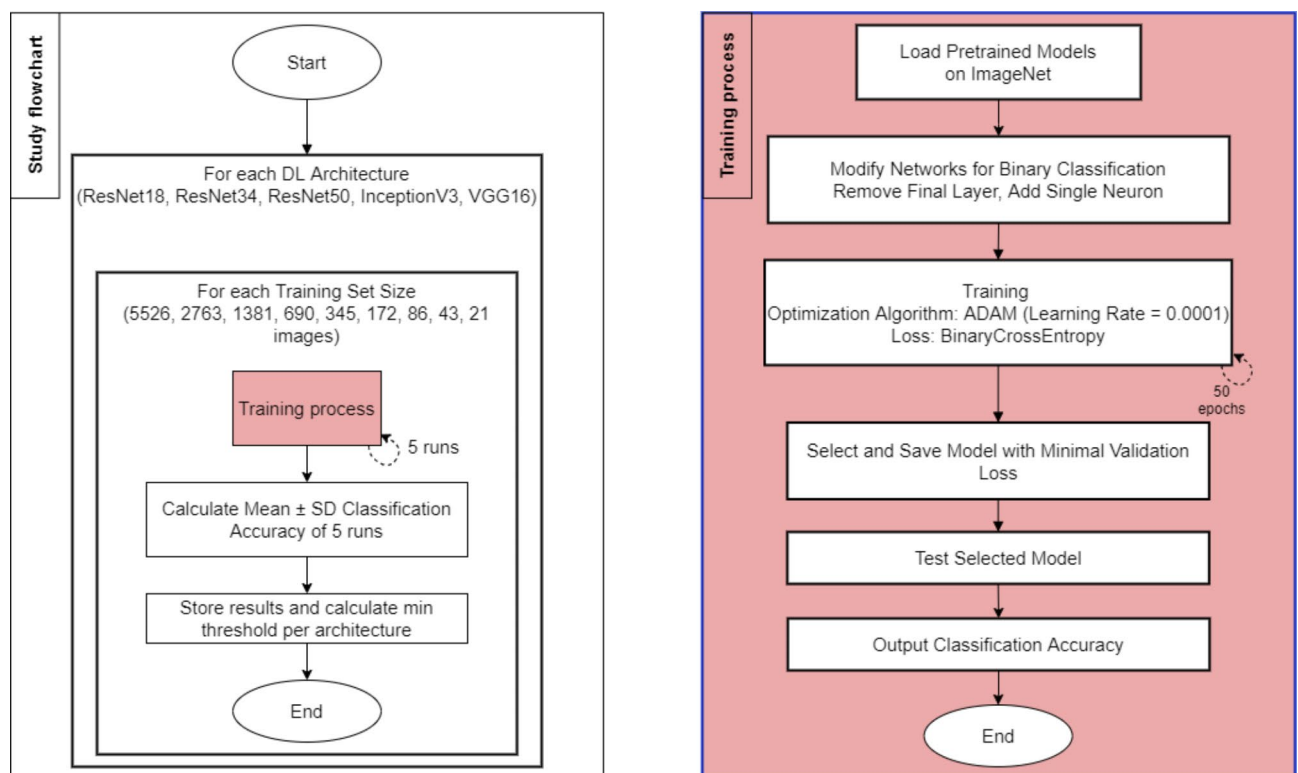


Fig. 2. Study flowchart for evaluating DL architectures (ResNet18, ResNet34, ResNet50, InceptionV3, VGG16) across various training set sizes.

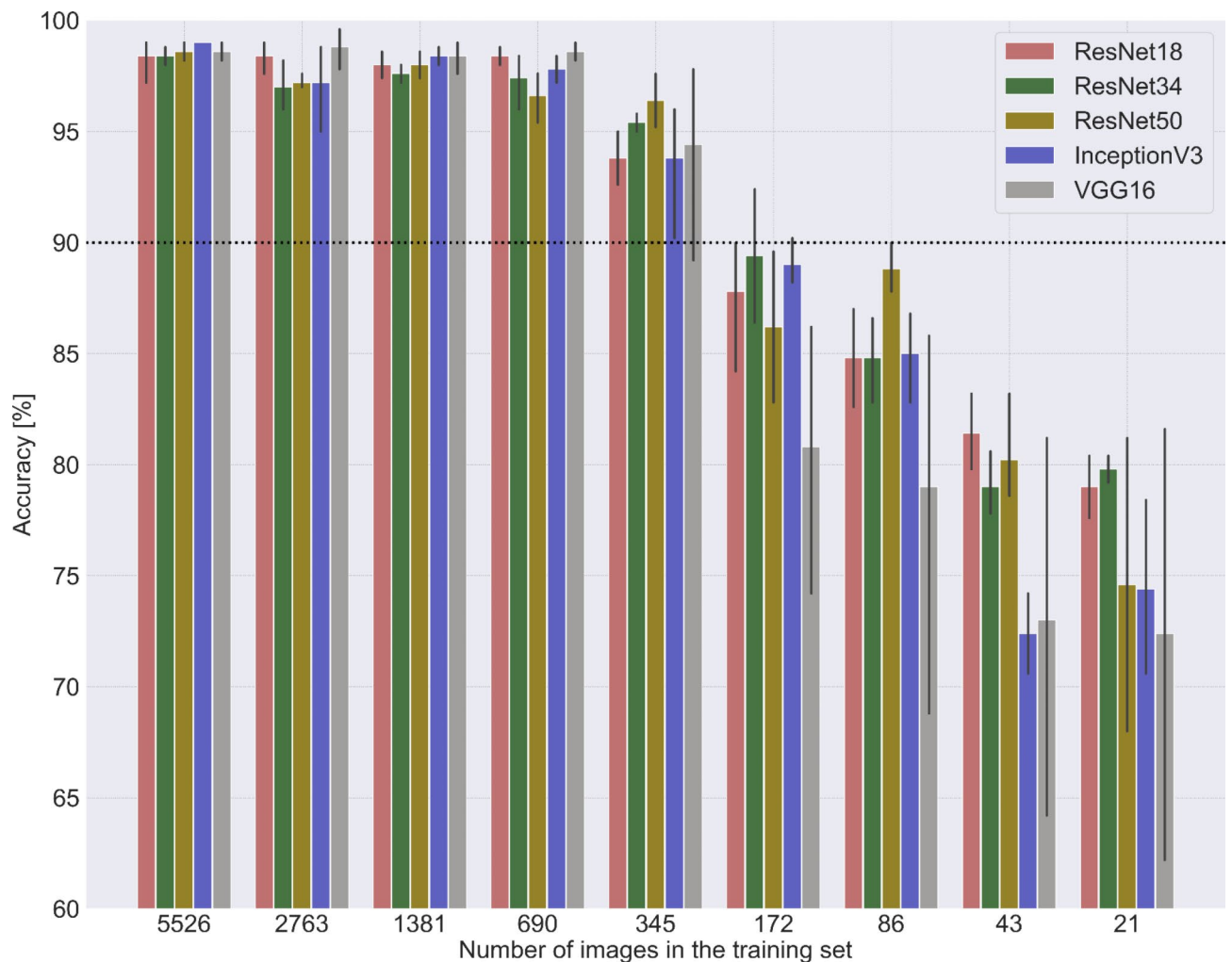


Fig. 3. The average (mean $\pm$ SD) accuracy of the models for each DL architecture, trained on different training subsets containing varying numbers of images. A horizontal black dotted line corresponds to the 90% classification accuracy.

For the smallest dataset sizes (43 and 21 images), accuracy drops significantly, with no model achieving the 90% threshold. ResNet34 and ResNet50 show more resilience to reductions in training set size compared to ResNet18, InceptionV3, and VGG16, with VGG16 exhibiting the steepest decline in performance and the largest standard errors at smaller dataset sizes. This increased variability in terms of SD underscores the instability and inconsistency of model performance under conditions of data scarcity.

Impact of noisy labeling

Figure 5 presents the test accuracies and confidence intervals (95% confidence level) for different DL architectures and training set sizes, in the presence of different levels of labeling noise (baseline - no mislabeling, 10%, 15%, and 20% mislabeling). For all DL architectures, as mislabeling is introduced, the accuracy of all models declines progressively with the label error rate. In addition, the impact of label noise becomes more pronounced with smaller training set sizes. However, some DL architectures exhibit slightly better robustness (Fig. 5).

In particular, when the training set was larger than 1381 images, ResNet34 and ResNet50 maintained classification performance for mislabeling up to 15%.

When the training set size is reduced to 690 and 345 images, the accuracy of all models notably drops with increasing levels of mislabeling. ResNet34 and ResNet50 also in this case showed better performance. For the mislabeling of 20% in all cases, the classification was considerably reduced, showing acceptable results in the presence of a high sample size.

Regarding variability in terms of standard deviation, non-misabeled models tend to maintain a similar interval of confidence width regardless of the number of images in the training set. In contrast, mislabeled models show greater variability as the number of images decreases. Finally, when these models are trained using the entire dataset of 5526 images, their performance variability is comparable to that of their non-misabeled versions.

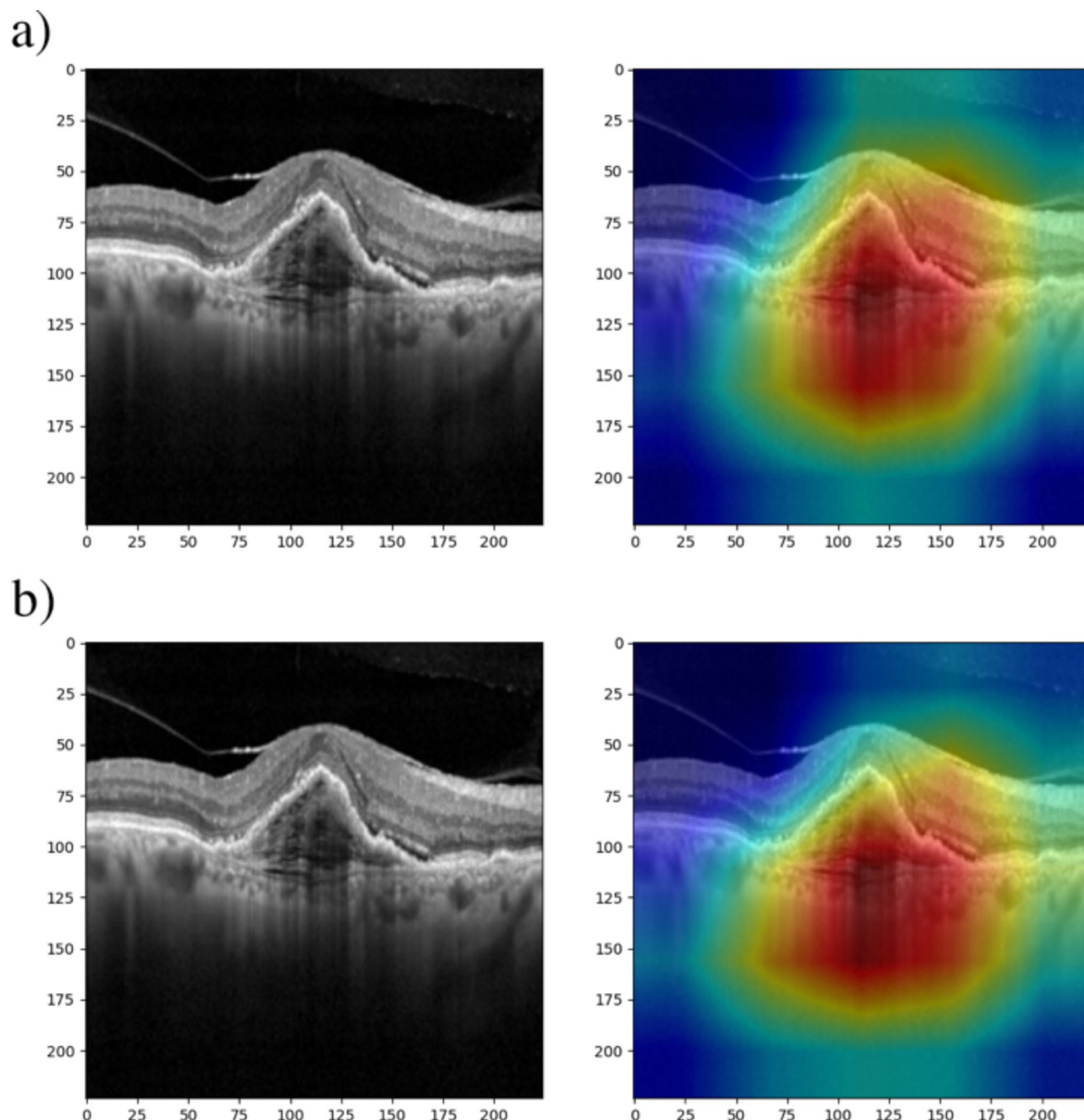


Fig. 4. Pathological OCT image and corresponding GradCam obtained from the model trained on **a)** 345 and **b)** 5526 images.

Considering the observed possibility to compensate for the accuracy reduction due to the miss-labeling errors with the increase in sample size, the analysis was conducted to assess the minimum sample size needed to achieve classification accuracy obtained on the non-mislabeled 345 images (Table 1). Among all architectures, VGG16 is the network that requires the fewest images to compensate for accuracy reduction due to labeling noise. Additional performance metrics, including sensitivity, specificity, and Matthews correlation coefficient, are available in the supplementary information of this paper (Table S1 – S16).

Discussion

The use of DL for OCT image classification has the potential to improve the diagnostic and monitoring procedures of retinal diseases²⁵. However, several challenges limit its clinical applicability³⁵. Key issues include the significant variability in retinal abnormalities, the presence of noise, and artifacts in OCT images. Thus large volumes of high-quality labeled data to train the models effectively are needed. However, the estimated number of images required to fine-tune these architectures and the impact of human labeling errors on model

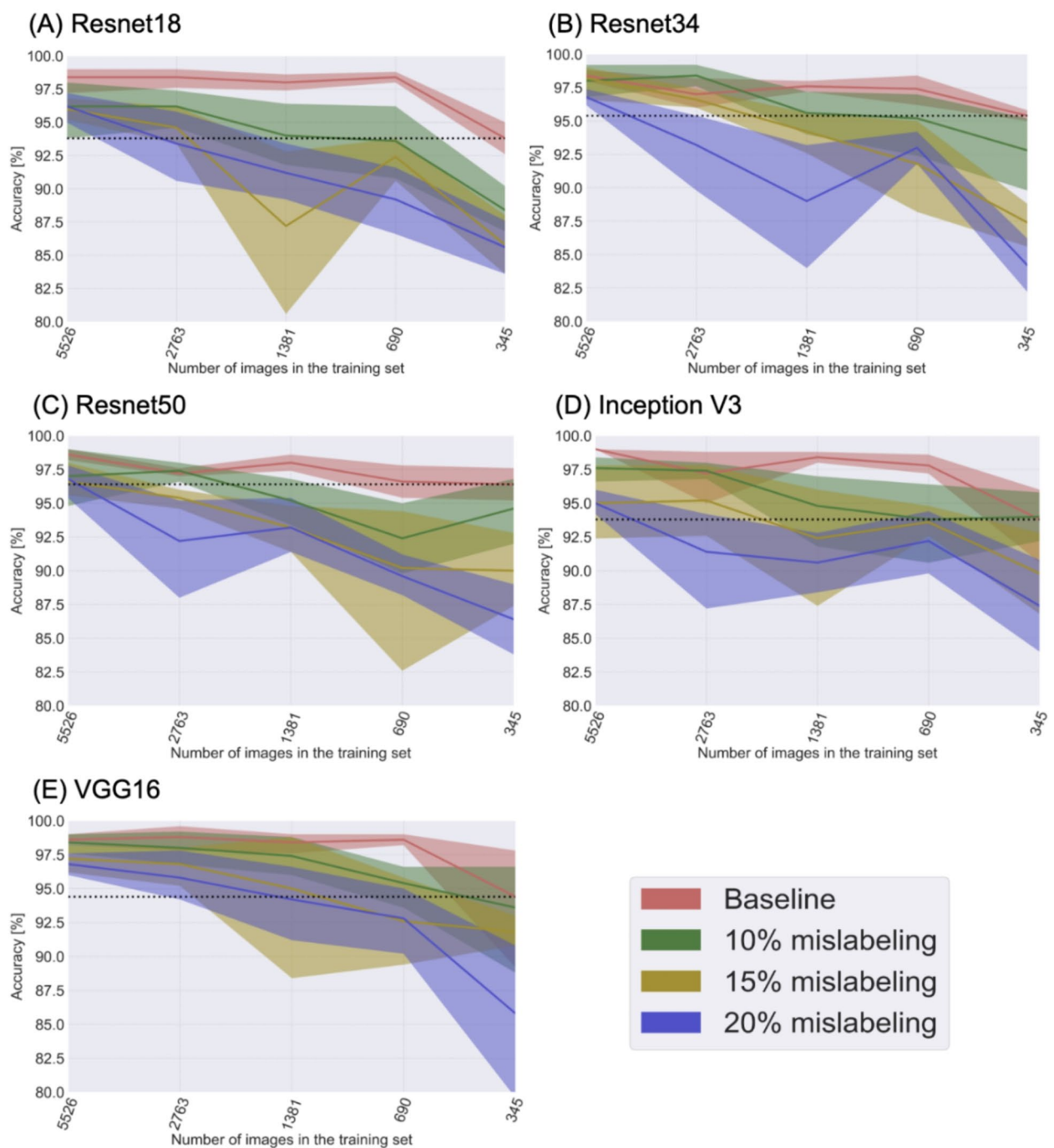


Fig. 5. Test accuracies and confidence intervals (95% confidence level) of various DL architectures with respect to the size of the training set. Each plot represents models developed with different levels of mislabeling in the training set: baseline (no mislabeling), 10%, 15%, and 20% mislabeling. The architectures evaluated are (A) ResNet18, (B) ResNet34, (C) ResNet50, (D) InceptionV3, and (E) VGG16. The horizontal black dotted line at 85% accuracy represents the clinically relevant threshold for model performance.

performance is an open issue. Furthermore, labeling OCT images is labor-intensive and requires expertise, often leading to human errors that introduce inconsistencies and biases, which can impact model performance²⁸ and limit clinical applicability.

Our study evaluated the performance of various DL architectures in classifying retinal pathologies versus healthy cases using OCT images under data scarcity and label noise.

Our results demonstrated that all considered architectures achieved high classification accuracy (>90%) when trained on datasets containing 345 or more images.

Architecture	10% Mislabeling	15% Mislabeling	20% Mislabeling
ResNet18	1381	2763	5526
ResNet34	1381	2763	5526
ResNet50	2763	5526	5526
InceptionV3	345	2763	5526
VGG16	690	1381	2763

Table 1. Number of images required for each DL architecture to achieve the performance of the baseline (no mislabeling) at 345 images, under different levels of mislabeling.

For larger sample sizes, no substantial differences in performance were observed among the different DL architectures, indicating that, given sufficient data, the choice of architecture may be less critical. However, InceptionV3 achieved the highest classification accuracy when trained on the entire dataset of 5526 images, suggesting a slight edge in performance.

Moreover, as the sample size decreased, there was a notable decline in classification accuracy and an increase in performance variability. All architectures maintained accuracy above 95% for dataset sizes down to 690 images, with minimal standard deviation, indicating consistent performance. These results align with findings from Rolnick et al. (2017)⁴⁷, who noted that DL models handle label noise better with larger datasets due to more robust feature learning.

Regarding the evaluation of the impact of label noise on the performance of various deep learning (DL) architectures for classifying retinal scans from OCT images. Our findings indicate that as mislabeling is introduced, the accuracy of all models declines and variability grows progressively with increasing label error rates. The impact of label noise becomes more pronounced as the training set size decreases. This is particularly evident for smaller datasets, where mislabeling significantly deteriorates model performance. However, some DL architectures have slightly better robustness to label noise. Among all the architectures, VGG16 demonstrates higher robustness in terms of the number of additional images required in the training set to reach the accuracy level of the model produced on correctly labeled 345 images. In the worst-case scenario of mislabeling, VGG16 is the network that requires the fewest images to adjust for this discrepancy.

Larger datasets provide more information, reducing the impact of noisy labels and enhancing model performance. This observation is supported by previous studies, which have shown that DL models are more robust to labeling noise with larger datasets⁵⁰.

Roughly estimated, across all architectures, on average 4, 9, and 14 times more images in the training set are required to compensate for labeling errors of 10%, 15%, and 20%, respectively. For most architectures, to mitigate the decrease in classification accuracy caused by 20% mislabeling, the entire dataset must be utilized.

This study has several limitations. The dataset, while substantial, may not encompass the full variability seen in clinical practice. The binary classification between normal and pathological OCT presented in this study may have limited direct application in clinical practice as a standalone decision support system. Additionally, the labeling errors were introduced uniformly, whereas real-world errors might exhibit more complex patterns. Furthermore, the study did not separately analyze the effects of different types of retinal diseases, which could provide more nuanced insights into model performance. However, the choice to combine different retinal abnormalities into a single pathological category was intended to increase the complexity of the classification problem on a clinically representative data sample. This approach required the model to recognize and differentiate a wide range of abnormal signs from normal OCT scans, adding a technical challenge beyond a simpler binary classification with a single pathology. We can therefore argue that these estimates are conservative and that in typical binary classification tasks (where the pathological category consists of a single condition), the models would likely perform even better than the results presented here.

Finally, these findings serve as a useful reference for clinical practice, particularly in the diagnosis and management of retinal diseases. By identifying the minimal data requirements and assessing the impact of label noise, this study offers the foundation for a framework for developing robust and reliable DL-based diagnostic tools, which can enhance the accuracy of detecting retinal pathologies at early stages and potentially be used as screening instruments.

Furthermore, the insights gained from this research can help integrate advanced DL models into routine ophthalmic care, thereby advancing precision medicine and improving patient outcomes.

Conclusions

In conclusion, our study systematically evaluated the performance of various deep learning (DL) architectures, including ResNet18, ResNet34, ResNet50, VGG16, and InceptionV3, in classifying retinal pathologies from OCT images under conditions of data scarcity and label noise. We found that all architectures achieved high classification accuracy (> 90%) when trained on 345 or more OCT images. Notably, InceptionV3 trained on the entire dataset exhibited the best performance.

However, a progressive decrease in average classification performance and an increase in variability were observed as the sample size decreased. Additionally, the presence of label noise significantly impacted model performance, showing more reduced accuracy and increased variability for higher mislabeling rates.

Across all architectures, compensating for labeling errors of 10%, 15%, and 20% requires approximately 4, 9, and 14 times more images in the training set to reach the performance obtained on 345 correctly labeled images. VGG16 demonstrated good resilience requiring fewer additional images to achieve comparable accuracy levels.

Our research provides a foundation for creating reliable DL-based diagnostic tools that can enhance the early detection and treatment of retinal diseases, inform clinical guidelines, and improve patient outcomes, thereby advancing the application of precision medicine in ophthalmology.

Data availability

The datasets generated and analyzed during the study are not publicly available due to privacy constraints. The data may however be available from the University of Trieste subject to local and national ethical approvals. In addition, we have made the models and relevant code available upon request. Any requests should be sent to the corresponding author.

Received: 3 September 2024; Accepted: 25 November 2024

Published online: 05 December 2024

References

- Duker, J. S., Waheed, N. K. & Goldman, D. R. Handbook of retinal OCT: Optical Coherence Tomography. *Elsevier Health Sci.* <https://doi.org/10.1016/C2015-0-00880-2> (2021).
- Zhang, L., Van Dijk, E. H. C., Borrelli, E., Fragiotta, S. & Breazzano, M. P. OCT and OCT angiography update: clinical application to age-related Macular Degeneration, Central Serous Chorioretinopathy, Macular Telangiectasia, and Diabetic Retinopathy. *Diagnostics* **13**, 232. <https://doi.org/10.3390/diagnostics13020232> (2023).
- Eladawi, N. et al. Classification of retinal diseases based on OCT images. *Front. Bioscience - Landmark*. **23**, 247–264. <https://doi.org/10.2741/4589> (2018).
- Özdaş, M. B., Uysal, F. & Hardalaç, F. Classification of Retinal diseases in Optical Coherence Tomography images using Artificial Intelligence and Firefly Algorithm. *Diagnostics* **13** <https://doi.org/10.3390/diagnostics13030433> (2023).
- Kadry, S., Rajinikanth, V., Taniar, D., Damaševičius, R. & Valencia, X. P. B. Automated segmentation of leukocyte from hematological images—a study using various CNN schemes. *J. Supercomput.* **78**, 6974–6994. <https://doi.org/10.1007/s11227-021-04125-4> (2022).
- Flores, R., Carneiro, A., Tenreiro, S. & Seabra, M. C. Retinal progression biomarkers of early and intermediate age-related macular degeneration. (2022). <https://doi.org/10.3390/life12010036>
- Bressler, N. M. Age-related macular degeneration is the leading cause of blindness. *JAMA* **291**, 1900–1901. <https://doi.org/10.1001/jama.291.15.1900> (2004).
- Schneider, E. W. & Fowler, S. C. Optical coherence tomography angiography in the management of age-related macular degeneration. *Curr. Opin. Ophthalmol.* **29**, 217–225. <https://doi.org/10.1097/ICU.0000000000000469> (2018).
- Corvi, F. et al. Optical coherence tomography angiography for detection of macular neovascularization associated with atrophy in age-related macular degeneration. *Graefes Archive Clin. Experimental Ophthalmol.* **259**, 291–299. <https://doi.org/10.1007/s00417-020-04821-6> (2021).
- Cheung, N., Mitchell, P. & Wong, T. Y. Diabetic retinopathy. *Lancet* **376**, 124–136. [https://doi.org/10.1016/S0140-6736\(09\)62124-3](https://doi.org/10.1016/S0140-6736(09)62124-3) (2010).
- Amoaku, W. M. et al. *Diabetic Retinopathy and Diabetic Macular Oedema Pathways and Management* (UK Consensus Working Group, 2020). <https://doi.org/10.1038/s41433-020-0961-6>
- Abbas, Q., Fondon, I., Sarmiento, A., Jiménez, S. & Alemany, P. Automatic recognition of severity level for diagnosis of diabetic retinopathy using deep visual features. *Med. Biol. Eng. Comput.* **55**, 1959–1974. <https://doi.org/10.1007/s11517-017-1638-6> (2017).
- Lindtjörn, B., Krohn, J. & Forsaa, V. A. Optical coherence tomography features and risk of macular hole formation in the fellow eye. *BMC Ophthalmol.* **21** <https://doi.org/10.1186/s12886-021-02111-1> (2021).
- Weinreb, R. N., Aung, T. & Medeiros, F. A. The pathophysiology and treatment of glaucoma: a review. *JAMA* **311**, 1901–1911. <https://doi.org/10.1001/jama.2014.3192> (2014).
- Geevarghese, A., Wollstein, G., Ishikawa, H. & Schuman, J. S. Optical coherence tomography and Glaucoma. *Annual Rev. Vis. Sci.* **7**, 693–726. <https://doi.org/10.1146/annurev-vision-100419-111350> (2021).
- Akpınar, M. H. et al. Artificial intelligence in retinal screening using OCT images: a review of the last decade (2013–2023). *Comput. Methods Programs Biomed.* **254**, 108253. <https://doi.org/10.1016/j.cmpb.2024.108253> (2024).
- Pandya, B. U., Grinton, M., Mandelcorn, E. D. & Felfeli, T. Retinal optical coherence tomography imaging biomarkers: a review of the literature. *Retina* **44**, 369–380. <https://doi.org/10.1097/IAE.0000000000003974> (2024).
- Kusuhara, S., Fukushima, Y., Ogura, S., Inoue, N. & Uemura, A. Pathophysiology of Diabetic Retinopathy: The Old and the New. *Diabetes Metab. J.* **42**, 364. <https://doi.org/10.4093/dmj.2018.0182> (2018).
- Algvere, P. V., Kvanta, A. & Seregård, S. Drusen maculopathy: a risk factor for visual deterioration. *Acta Ophthalmol.* **94**, 427–433. <https://doi.org/10.1111/aos.13011> (2016).
- Sheth, J. U. et al. Macular neovascularization. *Surv. Ophthalmol.* <https://doi.org/10.1016/j.survophthal.2024.08.003> (2024).
- Scheufele, T. A. et al. Photoreceptor atrophy. *Retina (Philadelphia Pa.)*. **25**, 1109. <https://doi.org/10.1097/00006982-200512000-00027> (2005). Acute Posterior Multifocal Placoid Pigment Epitheliopathy Demonstrated By Optical Coherence Tomography.
- Duker, J. S. et al. The International Vitreomacular Traction Study Group Classification of Vitreomacular Adhesion, Traction, and Macular Hole. *Ophthalmology* **120**, 2611–2619. <https://doi.org/10.1016/j.ophtha.2013.07.042> (2013).
- Bu, S. C., Kuijer, R., Li, X. R., Hooymans, J. M. M. & Los, L. I. Idiopathic Epiretinal Membrane *Retina* **34**, 2317 <https://doi.org/10.1097/IAE.0000000000000349>. (2014).
- Bikbova, G., Oshitari, T., Baba, T., Yamamoto, S. & Mori, K. Pathogenesis and management of Macular Hole: review of current advances. *J. Ophthalmol.* **3467381** (2019). <https://doi.org/10.1155/2019/3467381> (2019).
- LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444. <https://doi.org/10.1038/nature14539> (2015).
- Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88. <https://doi.org/10.1016/j.media.2017.07.005> (2017).
- Ronneberger, O. U-Net: Convolutional networks for biomedical image segmentation. In: [International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015).
- Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA - J. Am. Med. Association.* **316**, 2402–2410. <https://doi.org/10.1001/jama.2016.17216> (2016).
- Ting, D. S. W. et al. Development and validation of a Deep Learning System for Diabetic Retinopathy and Related Eye diseases using retinal images from multiethnic populations with diabetes. *JAMA* **318**, 2211–2223. <https://doi.org/10.1001/jama.2017.18152> (2017).
- Leandro, I. et al. OCT-based deep-learning models for the identification of retinal key signs. *Sci. Rep.* **13**, 14628. <https://doi.org/10.1038/s41598-023-41362-4> (2023).
- Ng, W. Y. et al. Updates in deep learning research in ophthalmology. *Clin. Sci. (Lond.)*. **135**, 2357–2376. <https://doi.org/10.1042/CS20210207> (2021).
- Moraru, A. D., Costin, D., Moraru, R. L. & Branisteanu, D. C. Artificial intelligence and deep learning in ophthalmology - present and future (review). *Experimental Therapeutic Med.* **20**, 3469–3473. <https://doi.org/10.3892/etm.2020.9118> (2020).

33. Straatsma, B. R. Precision medicine and clinical ophthalmology. *Indian J. Ophthalmol.* **66**, 1389. https://doi.org/10.4103/ijo.IJO_1459_18 (2018).
34. Schmidt-Erfurth, U. et al. Prediction of Individual Disease Conversion in Early AMD using Artificial Intelligence. *Investig. Ophthalmol. Vis. Sci.* **59**, 3199–3208. <https://doi.org/10.1167/iov.18-24106> (2018).
35. Yanagihara, R. T., Lee, C. S., Ting, D. S. W. & Lee, A. Y. Methodological challenges of Deep Learning in Optical Coherence Tomography for Retinal diseases: a review. *Translational Vis. Sci. Technol.* **9**, 11. <https://doi.org/10.1167/tvst.9.2.11> (2020).
36. Fujimoto, J. & Swanson, E. The Development, commercialization, and impact of Optical Coherence Tomography. *Investig. Ophthalmol. Vis. Sci.* **57**, OCT1–OCT13. <https://doi.org/10.1167/iov.16-19963> (2016).
37. Phua, D. H. & Tan, N. C. K. Cognitive aspect of diagnostic errors. *Ann. Acad. Med. Singap.* **42**, 33–41 (2013).
38. Payne, V. L. & Patel, V. L. Enhancing medical decision making when caring for the critically ill: the role of cognitive heuristics and biases. In: (eds Patel, V. L., Kaufman, D. R. & Cohen, T.) *Cognitive Informatics in Health and Biomedicine: Case Studies on Critical Care, Complexity and Errors*. 203–231. Springer, London https://doi.org/10.1007/978-1-4471-5490-7_10. (2014).
39. Feyzi-Behnagh, R. et al. Metacognitive scaffolds improve self-judgments of Accuracy in A Medical Intelligent Tutoring System. *Instr. Sci.* **42**, 159–181. <https://doi.org/10.1007/s11251-013-9275-4> (2014).
40. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. IEEE, Las Vegas, NV, USA (2016). <https://doi.org/10.1109/CVPR.2016.90>
41. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition, (2015). <http://arxiv.org/abs/1409.1556>.
42. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 2818–2826 (2016). <https://doi.org/10.1109/CVPR.2016.308>
43. Deng, J. et al. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). <https://doi.org/10.1109/CVPR.2009.5206848>
44. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (The MIT Press, 2016).
45. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization, (2017). <http://arxiv.org/abs/1412.6980>.
46. Janocha, K. & Czarnecki, W. M. On Loss Functions for Deep Neural Networks in Classification, (2017). <http://arxiv.org/abs/1702.05659>.
47. Rolnick, D., Veit, A., Belongie, S. & Shavit, N. Deep learning is robust to massive label noise. *arXiv Preprint arXiv:1705.10694*. <https://doi.org/10.48550/arXiv.1705.10694> (2017).
48. Algan, G. & Ulusoy, I. Image classification with deep learning in the presence of noisy labels: a survey. *Knowl. Based Syst.* **215**, 106771. <https://doi.org/10.1016/j.knosys.2021.106771> (2021).
49. Karimi, D., Dou, H., Warfield, S. K. & Gholipour, A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *Med. Image. Anal.* **65**, 101759. <https://doi.org/10.1016/j.media.2020.101759> (2020).
50. Frénay, B. & Verleysen, M. Classification in the presence of label noise: a survey. *IEEE Trans. Neural Networks Learn. Syst.* **25**, 845–869. <https://doi.org/10.1109/TNNLS.2013.2292894> (2014).

Author contributions

A.M. analyzed/interpreted data, wrote the article, proofed/revised article; A.B. analyzed/interpreted data, proofed/revised article; M.A. analyzed/interpreted data, proofed/revised article; S.K. analyzed/interpreted data, proofed/revised article; A.A. conducted the experiment, proofed/revised article; A.A. provided materials, analyzed/interpreted data, proofed/revised article; D.M. conducted the experiment, provided materials, proofed/revised article; T.D. provided materials, proofed/revised article. L.I. designed the experiment, conducted the experiment, provided materials, wrote the article; All authors reviewed the manuscript prior to submission.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-81127-1>.

Correspondence and requests for materials should be addressed to A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2024