A Trust-Based Methodology to Evaluate Deep Learning Models for Automatic Diagnosis of Ocular Toxoplasmosis from Fundus Images

Rodrigo Parra 1,*+, Verena Ojeda 1,†, Jose Luis Vázquez Noguera 1, Miguel García-Torres 2, Julio César Mello-Román 1, Cynthia Villalba 3, Jacques Facon 4, Federico Divina 2, Olivia Cardozo 5, Verónica Elisa Castillo 6 and Ingrid Castro Matto 6

Abstract: In the automatic diagnosis of ocular toxoplasmosis (OT), Deep Learning (DL) has arisen as a powerful and promising approach for diagnosis. However, despite the good performance of the models, decision rules should be interpretable to elicit trust from the medical community. Therefore, the development of an evaluation methodology to assess DL models based on interpretability methods is a challenging task that is necessary to extend the use of AI among clinicians. In this work, we propose a novel methodology to quantify the similarity between the decision rules used by a DL model and an ophthalmologist, based on the assumption that doctors are more likely to trust a prediction that was based on decision rules they can understand. Given an eye fundus image with OT, the proposed methodology compares the segmentation mask of OT lesions labeled by an ophthalmologist with the attribution matrix produced by interpretability methods. Furthermore, an open dataset that includes the eye fundus images and the segmentation masks is shared with the community. The proposal was tested on three different DL architectures. The results suggest that complex models tend to perform worse in terms of likelihood to be trusted while achieving better results in sensitivity and specificity.

Keywords: deep learning; ocular toxoplasmosis; machine learning interpretability; trust

1. Introduction

Over a third of the world’s human population is exposed to Toxoplasma gondii, making Toxoplasmosis one of the most common parasitic diseases worldwide [1]. Ocular toxoplasmosis (OT) occurs if the parasite reaches the retina, as it can damage host cells and neighboring cells leaving primary lesions. OT requires drug-based therapy to eliminate the parasite and the inflammation caused by it. If not treated properly, OT can lead to loss of vision [2].

Ophthalmologists conduct eye exams that look for lesions caused by the disease in eye fundus images to diagnose OT. Clinical manifestations of the disease tend to be highly characteristic; however, atypical manifestations can cause false-negative errors even by...
experienced doctors. Clinical examination is considered the diagnostic standard, due to the lack of a sufficiently sensitive lab test [3].

Machine learning is a subfield of artificial intelligence that allows computers to learn from existing data and make predictions. Its application has improved the performance of many challenging tasks in medical imaging, with a considerable impact on ophthalmology based on fundus photography, optical coherence tomography and slit-lamp imaging [4].

Deep learning (DL) is a subfield of machine learning based on artificial neural networks (ANN), a paradigm inspired by the human brain. DL models allow end-to-end learning, skipping the feature engineering step that was required by traditional computer vision approaches [5]. DL models have achieved promising results in automatic classification of images, and they have brought breakthroughs to the state of the art in recent years [6].

In particular, when applied to retinal images for medical diagnosis and prognosis, convolutional neural networks (CNNs) have been able to identify and estimate the severity of ocular diseases, such as age-related macular degeneration [7] and diabetic retinopathy [8,9]. Moreover, models have been trained to detect lesions caused by these diseases and classify them according to their severity [10].

Hasanreisoglu et al. [11] explored similar techniques for OT diagnosis using fundus images. Parra et al. [12] attempted an additional network architecture and achieved promising results, in addition to publishing an open OT dataset. To the best of our knowledge, these are the only works that have applied deep learning to OT diagnosis.

In the field, most works have been focused on the predictive power of the model. However, despite the good results obtained, the medical community is skeptical about its use due, mainly, to the difficulty in the interpretation of the results. Human factors play an important role in the diagnosis, and they must be taken into account to increase the reliability of the models induced and to extend human-AI collaboration. The concept of Trust arises in this context, defined as the intention to accept vulnerability based on positive expectations [13]. Currently, a lack of trust in AI systems is a significant drawback in the adoption of this technology in healthcare [14]. Understanding the reasons behind predictions, and analyzing them considering prior knowledge about the application domain, can be important to establish trust [15].

Zhang et al. defined interpretability as the ability to provide explanations in understandable terms to a human [16]. As such, interpretability methods can be used to obtain an explanation of the output of a predictive model. Attribution methods, a family of interpretability methods, assign credit (or blame) with regards to the prediction to the input features. For images, this means that they assign a score to each of the input pixels.

Several deep learning attribution methods are based on gradients, i.e., partial derivatives of the output with respect to the input. Gradient * Input [17], Integrated Gradients [18], Layer-wise Relevance Propagation (LRP) [19] and DeepLIFT [20] are examples of such methods. Although they use gradients differently to compute attribution scores, Ancona et al. have shown these methods to be strongly related, if not equivalent under certain conditions [21].

Attribution methods have been applied to classification problems with retinal images, to enrich predictions presented to physicians. Sayres et al. explored integrated gradients to grade diabetic retinopathy [22], and Mehta et al. used the same method for automatic detection of glaucoma [23].

A general-purpose trust metric was proposed by Wong et al. [24] and extended by Hryniewski et al. [25]. They were experimentally tested with Imagenet with insightful results. Interpretability, a prerequisite of trust, is known to be a domain-specific notion [26]. Hence, we argue that domain-specific trust metrics are important for machine learning adoption.

In this study, we propose a method to quantitatively evaluate the trustworthiness of a model in the OT diagnosis domain. We do this by comparing the average attribution scores of pixels that belong to a lesion vs. the rest of the pixels. We assume that doctors are
more likely to trust a model if its predictions are based on the features they consider for their diagnosis. Hence pixels within lesions should have higher attribution scores than the rest for an OT model to be considered trustworthy.

The rest of this paper is organized as follows. Section 2 introduces the main concepts of this work, including the data used. Then, in Section 3, the experimental results are described. The discussion about such results is given in Section 4. Finally, Section 5 presents the conclusions of this work.

2. Materials and Methods

In this section, the main characteristics of the data are first presented. Then, the different Deep Learning architectures are introduced. Finally, the proposed evaluation methods are described.

2.1. Dataset

Predictive models were trained and evaluated on a dataset of 160 eye fundus images. These images were collected at the Hospital de Clínicas in Asunción (Paraguay) by members of the Department of Ophthalmology. Some examples of the dataset can be seen in Figure 1. The complete dataset can be found online and is freely available for research purposes.

![Figure 1. A sample of healthy (a,b) and unhealthy (c,d) retinal fundus images from the dataset.](image)

Images were captured using a Zeiss brand camera, model Visucam 500, operated by experienced ophthalmologists. Each image was manually segmented by an ophthalmologist using an open source labeling tool ([https://labelstud.io](https://labelstud.io) (accessed on Wednesday, 20 October 2021)) to manually highlight OT entities (active lesions and inactive scars).

Active lesions have variable size, white or yellow color, blurry edges and a cottony center. They might be associated with a brown retinal hyperpigmentation area, which
is compatible with previous scar lesions. In some cases, active lesions can be hard to
differentiate due to the presence of vitreitis. Inactive lesions have variable size with
possible brown hyperpigmentation, with a stunted yellow or white center. An example of
these annotations can be seen in Figure 2.

![Figure 2. A sample of unhealthy eye fundus images (a,c) with their corresponding masks of segmented OT lesions (b,d) from the dataset.](image)

2.2. Model Training

Deep learning models and, in particular, CNNs, have achieved state-of-the-art results
in terms of predictive power for computer vision use cases [27]. Convolutional neural
networks are a particular type of feedforward neural networks (artificial neural networks
with no backlinks) that is normally composed of a combination of layers:

- Convolutional layers: capture local features by sliding a set of kernels over their input.
- Pooling layers: are used to downsample the output of convolutional layers.
- Fully-connected layers: are often used as the final layers of the model, to perform the
  final prediction.

As kernels share weights with all neurons, they help significantly in reducing the total
number of parameters of the network. Thus, CNNs allow building neural networks with
many layers with fewer parameters than other architectures [28].

We evaluate three different architectures:

- A CNN model with a few convolutional layers initialized with random weights.
- A VGG16 [29] model pretrained on the Imagenet dataset.
- A Resnet18 [30] model pretrained on the Imagenet dataset.
VGG16 is an architecture proposed by Simonyan and Zisserman, which was the first to experiment with smaller kernel sizes achieving promising results and increased depth of the model. Furthermore, Resnet18, which introduced the concept of residual connections. Residual connections help transfer knowledge from previous layers, alleviating the vanishing gradient problem that neural networks often suffer from. Residual networks allowed even deeper models to be trained, with a decreased number of parameters [28].

A comparison of the three architectures in terms of number of parameters and depth is shown in Table 1.

Table 1. A comparison of the three selected deep learning models.

| Model     | Parameters (Millions) | Layers |
|-----------|-----------------------|--------|
| Vanilla CNN | 5.6                   | 6      |
| VGG16     | 138                   | 16     |
| Resnet18  | 11                    | 152    |

Data augmentation based on random flips and crops was performed for all models, as shown in Figure 3. The last two models leverage transfer learning, i.e., they were pretrained on a larger general-purpose image dataset and then, with minor modifications to the learned weights, applied to OT classification for which less data is available. This is common when applying DL in domains where it is very difficult to build well-annotated datasets on a large scale due to the cost of acquiring data and annotations [31]. The idea of transfer learning is represented graphically in Figure 4.

Figure 3. An original eye fundus training image (a) with some example transformations, such as vertical flip (b), horizontal flip (c) and rotation (d), which are computed for data augmentation.
Models were optimized for 50 epochs using stochastic gradient descent (SGD) with a batch size of 32. Binary cross-entropy loss was used as the optimization target. The dataset was split into training (70%), validation (10%) and test (20%) sets. The training set was used for model fitting, the validation set for hyperparameter tuning and the test set to make the final model evaluation.

![Figure 4](image.png)

**Figure 4.** A schematic overview of transfer learning.

### 2.3. Model Evaluation

All models were evaluated using traditional predictive performance metrics: accuracy, sensitivity and specificity. In addition to that, we propose a method to obtain a trust score based on feature attributions, which is described in detail below. We only consider ed images with lesions that were correctly classified by the models (as a reminder, we consider an eye fundus image to be unhealthy if there are any lesions) for our evaluation, since our analysis depends on OT entities that were segmented by ophthalmologists.

#### 2.3.1. Measuring Feature Importance: Pixel Attribution Scores

Attribution methods provide scores for each of the input features that estimate the relevance they had on the prediction. Formally, given a deep neural network (DNN) $F : \mathbb{R}^n \rightarrow [0, 1]$, let $x \in \mathbb{R}^n$ be the model input. An attribution method can be seen as a function $A(F, x) = [s_1, \ldots, s_n]$, where $s_1, \ldots, s_n$ are referred to as attribution scores. In this study, we use Integrated Gradients (IG) as the attribution method of choice.

Let $x' \in \mathbb{R}^n$ be a baseline input of the model, which is usually a black image for image networks. Integrated gradients are defined as the integral of the gradients along the path from the baseline $x'$ to the input $x$. The integrated gradient for the $i$th dimension is defined as follows:

$$\text{IntegratedGradients}_i(x) := (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} \, d\alpha$$

where $\frac{\partial F(x)}{\partial x_i}$ is the gradient of $F(x)$ along the $i$th dimension and $\alpha$ is an interpolation constant to perturb features by.
We can calculate an attribution score per feature using IG. To obtain a per-pixel attribution score, we sum scores across RGB channels. The proposal of this study is independent of the actual attribution method selected.

2.3.2. Evaluating a Prediction: To Trust or Not to Trust?

Given a particular pixel attribution matrix \( A \in \mathbb{R}^n \) and a mask of OT entities for the original image, some pixels belong to an OT entity and others do not. Assume that those two groups of pixels were sampled from different populations, \( L \) and \( R \). We expect the median of \( L \) to be larger than that of \( R \) for OT cases, i.e., pixels from the lesions identified by a physician should be relatively more relevant for the model to elicit trust from them. We can test this hypothesis by using a one-sided Mann–Whitney U test such that:

- \( h_0 \): The median of \( R \) is larger or equal than the median of \( L \).
- \( h_1 \): The median of \( L \) is larger than the median of \( R \).

Therefore, we can define a binary trust function \( t \) as:

\[
t(A) = \begin{cases} 
0, & \text{if the } p\text{-value } < 0.05, \text{i.e., we fail to reject the null hypothesis} \\
1, & \text{otherwise}
\end{cases}
\]

2.3.3. Evaluating a Model Given a Dataset: Aggregating Our Results

Given a test set of images, a model is scored by calculating the ratio of images for which we obtain a one after applying \( t \) to their pixel-attribution matrix. This aggregate represents the proportion of images for which to model is likely to be considered trustworthy by an ophthalmologist.

The general purpose trust score proposed by Wong et al. [24] and extended by Hryniowski et al. [25] defines trust based on the answer to two questions: (1) How much trust do we have in a model that gives wrong answers with great confidence? and (2) How much trust do we have in a model that gives right answers hesitantly? However, valuable, interpretability and trust are known to be domain-specific notions [26]. Hence, the trust score proposed in this work incorporates domain-specific knowledge (masks) and compares it with the attribution matrix to answer the question: Did the model consider the features that an ophthalmologist would have taken into account (lesions) for this prediction?

A general overview of the process to evaluate a model is depicted in Figure 5 and can be summarized as follows: (i) an eye fundus dataset was collected by ophthalmologists at the Hospital de Clínicas of Asunción, Paraguay, (ii) physicians manually segmented OT entities for every image that had lesions, (iii) a predictive model is trained on a subset of the eye fundus dataset, (iv) pixel- attribution matrices are computed for all correctly-predicted sick images of a test set and, finally, (v) segmentation masks and attribution matrices are compared using a Mann–Whitney U test, and the results are aggregated to calculate the model trust score.
3. Results

The experiments were performed on a Google Colab Pro account, which provides Nvidia T4 and P100 graphic cards and up to 25 GB of RAM. The models were implemented using Pytorch 1.4. Models were trained with a batch size of 32, a learning rate of $1 \times 10^{-2}$ and stochastic gradient descent (SGD) as the optimizer, and these hyperparameters were selected according to the selection process performed by Parra et al. [12].

Two experiments were performed:

- Models were trained and evaluated with respect to accuracy, sensitivity and specificity, to contrast them with the results of the proposed trust metric and then,
- Models were evaluated using the proposed trust score on all correctly-predicted sick images from the test set.

Figure 5. A schematic overview of the general process for trust evaluation.
3.1. Common Predictive Metrics

After fine-tuning all predictive models common metrics used to evaluate predictive power were computed on the complete test set. Table 2 summarizes the results in terms of accuracy, sensitivity and specificity. The goal of this experiment was to determine if the evaluated models ranked similarly to comparisons made in other domains. As expected, both VGG and Resnet achieve better results than the vanilla CNN. Interestingly, better results were achieved with VGG than with Resnet as opposed to the results published for ImageNet [30].

Table 2. Predictive metrics comparison for the three deep learning models.

| Model   | Accuracy | Sensitivity | Specificity |
|---------|----------|-------------|-------------|
| Vanilla CNN | 0.75     | 0.75        | 0.75        |
| VGG16   | 0.96875  | 1.0         | 0.9375      |
| Resnet18| 0.9375   | 0.9375      | 0.9375      |

3.2. Trust Score

The proposed trust score was calculated for each of the models on the subset of correctly-labeled images from the test set, as depicted in Section 5. Aggregated results for each the compared models are summarized in Table 3. Predictive metrics are included to better contrast their relationship to the proposed score. The results show that models that scored higher in terms of traditional metrics associated with predictive power, e.g., accuracy, sensitivity and specificity performed worse in terms of the proposed trust score. This can be seen on a per-image basis in Figure 6. In addition to this, numeric values associated with the trust score calculation on a per-image basis can be found in Table A1 of Appendix A.

Table 3. Metric comparison including trust score for the three deep learning models.

| Model   | Accuracy | Sensitivity | Specificity | Trust  |
|---------|----------|-------------|-------------|--------|
| Vanilla CNN | 0.75     | 0.75        | 0.75        | 0.67   |
| VGG16   | 0.96875  | 1.0         | 0.9375      | 0.21   |
| Resnet18| 0.9375   | 0.9375      | 0.9375      | 0.14   |

Figure 6. Visualizing results: each cell represents an image of the test set that was predicted using a model. **Green** cells represent correct and trustworthy predictions (i.e., those where the lesions were relevant for the model output); **orange** cells are those where the model predicted the right label, but the prediction might not be trustworthy; and **red** cells are prediction errors.

4. Discussion

Exploratory analysis of the IG attribution maps confirms the intuition behind the proposed trust score. Figure 7 shows an example prediction for which the model was considered trustworthy. This can be visually verified as the attribution scores are clustered around the area of the lesion. Figure 8 shows an example prediction for which the model was considered untrustworthy. One can visually confirm that pixel attribution scores are scattered and less concentrated on the lesion area.
Figure 7. An example of an unhealthy eye fundus image that was correctly classified by the CNN model (a), the mask segmented by an ophthalmologist (b), a heatmap of the IG-based pixel attribution scores (c) and the attribution scores as an overlay (d). Median pixel attribution score differences were statistically significant between lesion and non-lesion areas.

Figure 8. An example of an unhealthy eye fundus image that was correctly classified by the Resnet18 model (a), the mask segmented by an ophthalmologist (b), a heatmap of the IG-based pixel attribution scores (c) and the attribution scores as an overlay (d). Median pixel attribution score differences were not statistically significant between lesion and non-lesion areas.

The obtained results suggest that predictions made by the most accurate deep learning might be harder to trust by experienced physicians. These findings agree with the existing literature, as it is known that healthcare workers often find it challenging to trust complex machine-learning models [32].
Interestingly, the relationship between the trust score and the number of parameters of the trained models (a common proxy for complexity) is not perfectly inverse. Although it is clear that the simple CNN scored much higher, the trust score for VGG16 was higher than that of Resnet18, despite having approximately 10-times more trainable parameters. This suggests that further research is needed regarding what exactly is it about complexity that punishes trustworthiness of the predictions, e.g., Are residual blocks bad for model trust? In other words, Can the key architectural decisions that lead to poor trustworthiness be identified?

Answering the previous question can lead to developing better building blocks for DL and machine learning in general, and this represents a needed, but challenging, change in the way state-of-the-art models are currently evaluated. Considering metrics beyond performance power is key to achieving mainstream adoption of predictive models in the healthcare domain.

5. Conclusions

We evaluated three different DL architectures and observed an inverse relation between the predictive power and our trust score. These results suggest that trust should also be considered for model selection, in addition to more traditional metrics, such as sensitivity and specificity. This is particularly the case if we expect deep learning models to be adopted by the medical community.

The main contributions of this work are: (i) an open dataset of annotated eye fundus images for OT diagnosis and (ii) a domain-specific method to evaluate predictive models with respect to trust (i.e., how likely a physician is to trust a model’s predictions) for OT diagnosis.

Extensions to our work can include: (i) a user study with ophthalmologists could help validate that our trust score adequately models their reactions to different model predictions, (ii) comparing the results using alternative attribution methods and (iii) comparing our score with traditional ML models by using an extension of IG that supports non-differentiable models [33].

Author Contributions: Conceptualization, methodology and formal analysis, R.P., V.O. and J.L.V.N.; data curation, J.C.M.-R., O.C., V.E.C. and I.C.M.; software, V.O.; writing—original draft, R.P.; writing—review and editing, M.G.-T., C.V., J.F. and F.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Consejo Nacional de Ciencia y Tecnología (CONACYT) as a part of project PINV18-1293.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Institutional Review Board of the Hospital Pediátrico “Niños de Acosta Nu” with the approval number 00197 on May, 2019.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data used to conduct this work is available at https://doi.org/10.5281/zenodo.4479724 (accessed on Wednesday, 20 October 2021).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

| Abbreviation | Description                      |
|--------------|----------------------------------|
| AI           | Artificial Intelligence          |
| DL           | Deep Learning                    |
| OT           | Ocular Toxoplamosis              |
| CNN          | Convolutional Neural Network      |
| IG           | Integrated Gradients             |
Appendix A. Detailed Results

Table A1. Results per pair of model and test image. Images are identified by a unique code (IID). Results include the mean attribution score for lesion-related and non-lesion-related pixels and the p-value obtained for the Mann–Whitney test.\(^1\)

| IID | Vanilla CNN | VGG16 | Resnet18 |
|-----|-------------|-------|----------|
|     | Mean lesion attribution: 0.0003472176724766176 | Mean lesion attribution: 0.0023451966569500915 | Mean lesion attribution: -6.291 439 449 744 539 \times 10^{-5} |
| 81  | Mean non-lesion attribution: -6.601 419 281 104.99 \times 10^{-6} | Mean non-lesion attribution: -1.505 774 503 219 224.6 \times 10^{-5} | Mean non-lesion attribution: 4.173 565 166 532.544 \times 10^{-6} |
|     | Mann–Whitney p-value: 0.0 | Mann–Whitney p-value: 2.182 831 996 769.315 \times 10^{-7} | Mann–Whitney p-value: 0.8461449179567923 |
|     | Mean lesion attribution: 7.608 095 877 901 965 \times 10^{-5} | Mean lesion attribution: 9.139 120 141 825 367 \times 10^{-6} | Mean lesion attribution: 1.026 627 135 195 607 \times 10^{-5} |
|     | Mean non-lesion attribution: 0.0004341474766965302 | Mean non-lesion attribution: 0.12678531658770215 | Mean non-lesion attribution: 1.283 862 080 623 880.5 \times 10^{-5} |
|     | Mann–Whitney p-value: 0.0 | Mann–Whitney p-value: 0.93552513605340 | Mann–Whitney p-value: 0.5119256881011157 |
| 156 | misclassified | misclassified | misclassified |
| 144 | Mean lesion attribution: 6.443 540 099 421 086 \times 10^{-5} | Mean lesion attribution: 9.005 965 358 771 226 \times 10^{-5} | Mean lesion attribution: 2.337 102 520 230 451 \times 10^{-5} |
|     | Mean non-lesion attribution: -3.871 953 049 444 241 \times 10^{-6} | Mean non-lesion attribution: -2.815 535 276 382 608.5 \times 10^{-5} | Mean non-lesion attribution: 1.267 964 040 454 379 \times 10^{-5} |
|     | Mann–Whitney p-value: 0.0 | Mann–Whitney p-value: 0.12678531658770215 | Mann–Whitney p-value: 0.93552513605340 |
| 97  | Mean lesion attribution: 6.336 514 502 731 98 \times 10^{-5} | Mean lesion attribution: 9.000 314 315 215 689 153 | Mean lesion attribution: 8.424 714 848 856 002 \times 10^{-6} |
|     | Mean non-lesion attribution: 1.029 527 369 847 149 3 \times 10^{-6} | Mean non-lesion attribution: -9.295 814 133 027 072 \times 10^{-7} | Mean non-lesion attribution: 1.200 708 985 033 684 \times 10^{-5} |
|     | Mann–Whitney p-value: 0.0 | Mann–Whitney p-value: 0.2580491242701976 | Mann–Whitney p-value: 0.976338912018821 |
| 151 | Mean lesion attribution: 8.273 648 382 325 341 \times 10^{-5} | Mean lesion attribution: 4.000 040 678 326 502 \times 10^{-6} | Mean lesion attribution: 5.638 891 517 729 174 \times 10^{-5} |
|     | Mean non-lesion attribution: -1.586 017 485 948 515 7 \times 10^{-6} | Mean non-lesion attribution: -4.000 040 678 326 502 \times 10^{-6} | Mean non-lesion attribution: 6.625 146 340 744 238 \times 10^{-6} |
|     | Mann–Whitney p-value: 0.0 | Mann–Whitney p-value: 0.03181811783986231 | Mann–Whitney p-value: 0.4084209761515357 |
| 142 | Mean lesion attribution: 9.391 440 979 266 882 \times 10^{-5} | Mean lesion attribution: 4.000 040 678 326 502 \times 10^{-6} | Mean lesion attribution: 0.0001558430584637686 |
|     | Mean non-lesion attribution: 5.027 028 776 625 433 6 \times 10^{-6} | Mean non-lesion attribution: 1.188 879 169 016 227 2 \times 10^{-5} | Mean non-lesion attribution: 6.200 042 851 696 207 \times 10^{-6} |
|     | Mann–Whitney p-value: 0.0 | Mann–Whitney p-value: 0.1151630357912255 | Mann–Whitney p-value: 0.3019700368885526 |
| 118 | Mean lesion attribution: -2.179 997 111 096 857 7 \times 10^{-5} | Mean lesion attribution: 7.862 114 094 452 912 \times 10^{-6} | Mean lesion attribution: 9.424 270 874 506 554 |
|     | Mean non-lesion attribution: 1.265 590 910 444 417 7 \times 10^{-5} | Mann–Whitney p-value: 0.7708368497428252 | Mann–Whitney p-value: 0.696399305035022 |

\(^1\)Statistical significance is determined using the Mann–Whitney U test with a significance level of p < 0.05.
### Table A1. Cont.

| IID | Vanilla CNN | VGG16 | ResNet18 |
|-----|-------------|-------|----------|
| 150 | Mean lesion attribution: −7.918 849 021 031 336 × 10⁻⁵<br>Mean non-lesion attribution: 1.583 111 698 817 516 7 × 10⁻⁵<br>Mann–Whitney p-value: 1.0 | Mean lesion attribution: 0.00013991786062732565<br>Mean non-lesion attribution: 9.315 094 828 690 027 × 10⁻⁶<br>Mann–Whitney p-value: 0.6484420024657709 | Mean lesion attribution: 1.135 622 422 691 986 × 10⁻⁵<br>Mean non-lesion attribution: 1.006 649 973 543 553 3 × 10⁻⁵<br>Mann–Whitney p-value: 0.5955185797651998 |
| 94  | misclassified | Mean lesion attribution: 0.00046278512907457894<br>Mean non-lesion attribution: −6.857 434 254 485 007 × 10⁻⁶<br>Mann–Whitney p-value: 1.687 138 254 138 508 6 × 10⁻⁹ | Mean lesion attribution: 2.31910601366875e-05<br>Mean non-lesion attribution: 1.304 558 028 864 510 6 × 10⁻⁶<br>Mann–Whitney p-value: 0.7233509758415473 |
| 132 | misclassified | Mean lesion attribution: 0.0007116612771495348<br>Mean non-lesion attribution: −5.007 881 843 520 115 × 10⁻⁶<br>Mann–Whitney p-value: 0.9661658611357002 | Mean lesion attribution: 3.984 074 887 705 268 × 10⁻⁶<br>Mean non-lesion attribution: 1.844 952 743 705 595 8 × 10⁻⁷<br>Mann–Whitney p-value: 0.9659463877638184 |
| 146 | Mean lesion attribution: 6.460267789192614e-05<br>Mean non-lesion attribution: 5.149 698 277 781 459 5 × 10⁻⁶<br>Mann–Whitney p-value: 3.629 556 224 709 794 × 10⁻¹⁵ | Mean lesion attribution: 0.00033481917634950995<br>Mean non-lesion attribution: −5.838 907 669 173 114 × 10⁻⁶<br>Mann–Whitney p-value: 0.9474421613415959 | Mean lesion attribution: 3.984 074 887 705 268 × 10⁻⁶<br>Mean non-lesion attribution: 1.844 952 743 705 595 8 × 10⁻⁷<br>Mann–Whitney p-value: 0.888148954079111 |
| 99  | Mean lesion attribution: 1.062346801897256e-05<br>Mean non-lesion attribution: −8.381 453 520 194 11 × 10⁻⁷<br>Mann–Whitney p-value: 8.705 342 309 642 405 × 10⁻¹² | Mean lesion attribution: 0.0001176112771495348<br>Mean non-lesion attribution: 1.122 893 524 636 988 8 × 10⁻⁶<br>Mann–Whitney p-value: 0.9991945637023955 | Mean lesion attribution: 4.387 994 697 815 6 × 10⁻⁵<br>Mean non-lesion attribution: 5.952 850 761 770 384 × 10⁻⁶<br>Mann–Whitney p-value: 0.888148954079111 |
| 117 | Mean lesion attribution: −3.639 901 875 997 307 6 × 10⁻⁵<br>Mean non-lesion attribution: 1.448 736 389 989 063 7 × 10⁻⁵<br>Mann–Whitney p-value: 1.0 | Mean lesion attribution: 0.0001403303076260771<br>Mean non-lesion attribution: 1.199 786 422 310 879 6 × 10⁻⁵<br>Mann–Whitney p-value: 0.8978942267882748 | Mean lesion attribution: 6.225647710709875e-06<br>Mean non-lesion attribution: 1.240 475 023 789 905 8 × 10⁻⁶<br>Mann–Whitney p-value: 0.503937509947166 |
| 111 | Mean lesion attribution: 6.573 892 462 999 344 × 10⁻¹⁻⁵<br>Mean non-lesion attribution: 6.415 230 420 503 763 × 10⁻⁶<br>Mann–Whitney p-value: 0.9986524555283746 | misclassified | Mean lesion attribution: 0.00011425205913621127<br>Mean non-lesion attribution: −2.209 874 994 196 942 × 10⁻⁶<br>Mann–Whitney p-value: 0.024341820096032026 |

1. *p*-values of 0.0 are smaller than the smallest floating point number representable by Numpy.

### References

1. Tenter, A.M.; Heckereth, A.R.; Weiss, L.M. Toxoplasma gondii: from animals to humans. *Int. J. Parasitol.* **2000**, *30*, 1217–1258. [CrossRef]
2. Park, Y.H.; Nam, H.W. Clinical features and treatment of ocular toxoplasmosis. *Korean J. Parasitol.* **2013**, *51*, 393–399. [CrossRef] [PubMed]
3. Garweg, J.G.; de Groot-Mijnes, J.D.F.; Montoya, J.G. Diagnostic approach to ocular toxoplasmosis. *Ocul. Immunol. Inflamm.* **2011**, *19*, 255–261. [CrossRef] [PubMed]
4. Tong, Y.; Lu, W.; Yu, Y.; Shen, Y. Application of machine learning in ophthalmic imaging modalities. *Eye Vis.* **2020**, *7*, 22. [CrossRef] [PubMed]
5. Mahony, N.O.; Campbell, S.; Carvalho, A.; Harapanahalli, S.; Velasco-Hernandez, G.; Krpalkova, L.; Riordan, D.; Walsh, J. Deep Learning vs. Traditional Computer Vision. In Proceedings of the Science and Information Conference, Las Vegas, NV, USA, 25–26 April 2019.

6. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. Nature 2015, 521, 436–444. [CrossRef]

7. Pead, E.; Megaw, R.; Cameron, J.; Fleming, A.; Dhillon, B.; Trucco, E.; MacGillivray, T. Automated detection of age-related macular degeneration in color fundus photography: A systematic review. Surv. Ophthalmol. 2019, 64, 498–511. [CrossRef]

8. Alyoubi, W.L.; Shalash, W.M.; Abulkhair, M.F. Diabetic retinopathy detection through deep learning techniques: A review. Inform. Med. Unlocked 2020, 20, 103377. [CrossRef]

9. Tsiknakis, N.; Theodoropoulos, D.; Manikis, G.; Ktistakis, E.; Boutsora, O.; Berto, A.; Scarpa, F.; Scarpa, A.; Fotiadis, D.I.; Marias, K. Deep learning for diabetic retinopathy detection and classification based on fundus images: A review. Comput. Biol. Med. 2021, 135, 104599. [CrossRef] [PubMed]

10. Yang, Y.; Li, T.; Li, W.; Zhang, W. Lesion Detection and Grading of Diabetic Retinopathy via Two-Stage Deep Convolutional Neural Networks. In International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Berlin/Heidelberg, Germany, 2017; pp. 533–540.

11. Hasanreisoglu, M.; Halim, M.S.; Chakravarthy, A.D.; Ormaecha, M.S.; Uludag, G.; Hassan, M.; Ozdemir, H.B.; Ozdal, P.C.; Colombo, D.; Rudzinski, M.N.; et al. Ocular Toxoplasmosis Lesion Detection on Fundus Photograph using a Deep Learning Model. Invest. Ophthalmol. Vis. Sci. 2020, 61, 1627.

12. Parra, R.; Ojeda, V.; Vázquez Noguer, J.L.; García Torres, M.; Mello Román, J.C.; Villalba, C.; Facon, J.; Divina, F.; Cardozo, O.; Castro, V.E.; et al. Automatic Diagnosis of Ocular Toxoplasmosis from Fundus Images with Residual Neural Networks. Stud. Health Technol. Inform. 2021, 281, 173–177.

13. Lockey, S.; Gillespie, N.; Holm, D.; Someh, I.A. A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. In Proceedings of the 54th Hawaii International Conference on System Sciences, Kauai, HI, USA, 5–8 January 2021.

14. Asan, O.; Bayrak, A.E.; Choudhury, A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. J. Med. Internet Res. 2020, 22, e15154. [CrossRef]

15. Ribeiro, M.T.; Singh, S.; Guestrin, C. “How Should I Trust You?”: Explaining the Predictions of Any Classifier. arXiv 2016, arXiv:1602.04938.

16. Zhang, Y.; Tiño, P.; Leonardis, A.; Tang, K. A Survey on Neural Network Interpretability. IEEE Trans. Emerg. Top. Comput. Intell. 2021, 5, 726–742. [CrossRef]

17. Shrikumar, A.; Greenside, P.; Shcherbina, A.; Kundaje, A. Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. arXiv 2016, arXiv:1605.01713.

18. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.

19. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. PLoS ONE 2015, 10, e0130140. [CrossRef] [PubMed]

20. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.

21. Ancona, M.; Ceolini, E.; Oztireli, C.; Gross, M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. arXiv 2017, arXiv:1711.06104.

22. Sayres, R.; Taly, A.; Rahimy, E.; Blumer, K.; Sohr, D.; Hammel, N.; Krause, J.; Narayanaswamy, A.; Rastegar, Z.; Wu, D.; et al. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. Ophthalmology 2019, 126, 552–564. [CrossRef]

23. Mehta, P.; Lee, A.Y.; Wen, J.; Bannit, M.R.; Chen, P.P.; Bojikian, K.D.; Petersen, C.; Egan, C.A.; Lee, S.I.; Balazinska, M.; et al. Automated detection of glaucoma using retinal images with interpretable deep learning. Invest. Ophthalmol. Vis. Sci. 2020, 61, 1150.

24. Wong, A.; Wang, X.Y.; Hryniowski, A. How Much Can We Really Trust You? Towards Simple, Interpretable Trust Quantification Metrics for Deep Neural Networks. arXiv 2020, arXiv:2009.05835.

25. Hryniowski, A.; Wang, X.Y.; Wong, A. Where Does Trust Break Down? A Quantitative Trust Analysis of Deep Neural Networks via Trust Matrix and Conditional Trust Densities. arXiv 2020, arXiv:2009.14701.

26. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. Electronics 2019, 8, 382. [CrossRef]

27. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. Comput. Intell. Neurosci. 2018, 2018, 7608349. [CrossRef]

28. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaria, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J. Big Data 2021, 8, 53. [CrossRef] [PubMed]

29. Simonian, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 2014, arXiv:1409.1556.

30. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
31. Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; Liu, C. A Survey on Deep Transfer Learning. In Proceedings of the International Conference on Artificial Neural Networks, Rhodes, Greece, 4–7 October 2018.
32. Stiglic, G.; Kocbek, P.; Fijacko, N.; Zitnik, M.; Verbert, K.; Cilar, L. Interpretability of machine learning based prediction models in healthcare. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2020, 10, e1379. [CrossRef]
33. Merrill, J.; Ward, G.; Kamkar, S.; Budzik, J.; Merrill, D. Generalized Integrated Gradients: A practical method for explaining diverse ensembles. arXiv 2019, arXiv:1909.01869.