IMPROVING EXPLANATIONS OF IMAGE CLASSIFIERS: ENSEMBLES AND MULTITASK LEARNING

Michael Pazzani[1], Severine Soltani[2,] Sateesh Kumar[3]
Kamran Alipour[3], and Aadil Ahamed[3]

[1]Information Sciences Institute, Marina Del Rey, CA, USA
[2]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA
[3]Department of Computer Science and Engineering,
University of California, San Diego, La Jolla, CA, USA

ABSTRACT

In explainable AI (XAI) for deep learning, saliency maps, heatmaps, or attention maps are commonly used to identify important regions for the classification of images of explanations. We address two important limitations of heatmaps. First, they do not correspond to type of explanations typically produced by human experts. Second, recent research has shown that many common XAI methods do not accurately identify the regions that human experts consider important. We propose using multitask learning to identify diagnostic features in images and averaging explanations from ensembles of learners to increase the accuracy of explanations. Our technique is general and can be used with multiple deep learning architectures and multiple XAI algorithms. We show that this method decreases the difference between regions of interest of XAI algorithms and those identified by human experts and the multitask learning supports the type of explanations produced by human experts. Furthermore, we show that human experts prefer the explanations produced by ensembles to those of individual networks.

KEYWORDS

Neural Networks, Machine Learning, Explainable AI, Image Classification, Computer Vision

1. INTRODUCTION

A variety of eXplainable Artificial Intelligence (XAI) methods have emerged for explaining image classification [16, 19] to developers or end-users [17, 5]. These approaches typically locate and highlight regions of the image that are important to the classification decision. Recently, several papers have called into question the ability of existing XAI methods to accurately identify regions that are meaningful to human experts such as radiologists, dermatologists, neurologists, oncologists, ophthalmologists, or even bird watchers [32, 9, 35, 24]. For example, there are substantial differences between the regions on an x-ray that radiologists find important and those found by XAI algorithms [2]. Furthermore, the dominant method for explaining image classification is assigning an importance score to pixels or regions on a saliency map or heatmap superimposed on an image, visualizing a region's importance with color scales (red, orange, yellow…). Although heatmaps unquestionably provide useful information to developers and perhaps technical auditors, particularly to indicate when the classifier mistakenly focuses on irrelevant regions of images, we argue they do not match what experts naturally produce nor what users expect.

DOI: 10.5121/ijaia.2022.13604

There are multiple purposes for XAI. One is to inform developers and perhaps validators how the deep learning system is working. Figure 1 shows images from XAI and from experts for image classification.

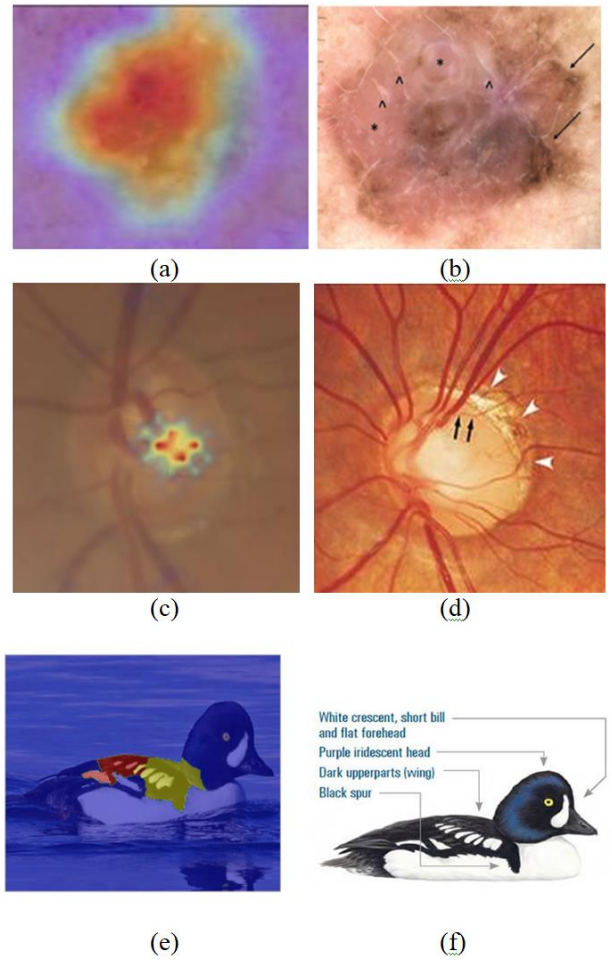


Figure 1. (a) Heatmap for explaining melanoma classification. (b) Image from medical journal explaining melanoma diagnosis. (c) Heatmap explaining glaucoma diagnosis (d) Image from medical journal explaining glaucoma diagnosis (e) Heatmap for explaining bird classification. (f) Image from a birding web site explaining bird classification

The left column shows example heatmaps generated in our lab for melanoma classification, glaucoma classification and bird species classification. In contrast to the left column of Figure 1, the right column shows explanations produced by experts to communicate with others. Figure 2b explains a melanoma diagnosis [22] with three regions identified and labeled with “milky pink structureless areas centrally (*), white streaks (^) and atypical pigment network (arrows). Figure 2d describes an unusual glaucoma case [16] indicating “parapapillary atrophy (arrowheads) and rim notching (arrows).” Figure 2f [36] also labels regions of the bird to explain its identifications. After many years of reading medical journals and bird guides, we have yet to encounter a heatmap used to explain image classification except in the developer-oriented context of describing deep learning.

There are two key differences between current XAI methods and the expert explanations.

- The most important difference is that the regions are labeled with semantically meaningful features. Terms such as "white streaks" in melanoma, rim notching in glaucoma, or short bill in bird identification have semantic meaning to experts and can be taught to novices. Rather than simply indicating all regions that led to a classification, experts have different labels for different regions. Instead of mapping images directly to conclusions, experts have conclusions about intermediate findings or diagnostic features.
- Expert explanations often use arrows to indicate important regions rather than using arbitrary polygons. These arrows are labeled with the semantically meaningful features.

In this paper, we will describe a new method to label images in the format that experts use. We describe deep learning architectures that are designed with the goals of accurate classifying images and identifying the diagnostic features important to this classification. Furthermore, we will adapt existing XAI algorithms to find the regions important to the classification. We use multitask learning [6] to simultaneously train the network on a class label and whether the image contains each potentially diagnostic feature.

We explore the use of ensemble learning [9] of neural networks to increase the accuracy of identifying regions of interest for any XAI algorithm by combining explanations from multiple neural networks. To illustrate, Figure 2(a-d) shows the heatmap of four neural networks trained on the same image data starting with different initial random weights. The task was to determine the wing pattern (e.g., striped, solid, spotted, wingbar). Figure 3(e) shows the heatmap produced by averaging the heatmaps of 11 networks. Of course, a disadvantage of our approach is that it requires more computation to create an ensemble than a single network. This linear increase in computation can be mitigated by coarse-grained parallel training of N networks.

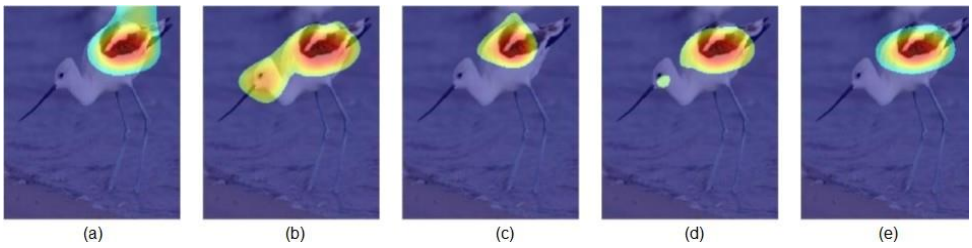


Figure 2. Saliency maps from an ensemble of classifiers. (a-d) are individual networks trained to identify the wing pattern. (e) is an average of 11 networks.

In the remainder of this paper, we first describe the methods we use to generate an ensemble of networks. Second, we discuss our evaluation methods which compare the regions of interest of an XAI algorithm to the regions of interest identified by people. Third, we describe the databases used in evaluation. Fourth, we describe the results using ensembles starting with different random weights on several problems. Fifth, we generalize our results by using two additional approaches to generating an ensemble of networks. Furthermore, we show the impact of varying the number of networks in the ensemble. Next, we present results from an experiment with experienced bird watchers that show that they prefer the explanations from ensembles to those of individual networks and that they prefer explanations with semantically labelled features as shown on the right of Figure 1. Finally, we show that multitask learning can create the explanations preferred by people.

2. PRIOR WORK

There are two main approaches in XAI to identify regions of interest in deep learning for image classification:

- Model agnostic methods, such as LIME [25], manipulate inputs (e.g., pixels, regions or more commonly superpixels) and measure how changes in input affect output. If an input perturbation has no effect, it is not relevant to findings. If a change has a major impact (e.g., changing the classification from pneumonia to normal), then the region is important to the classification. Shapley Additive Explanations (SHAP) [22] uses a game-theoretic measure to assign each feature or region an importance value for a particular prediction.
- Other methods examine the activations or weights of the deep network to find regions of importance. Grad-CAM [26], Integrated Gradients [33], Saliency [30], GradientShap [21], and Layerwise Relevance Propagation LRP [27] are examples of such methods.

Ensemble learning has long been used to reduce the error of machine learning methods, including neural networks [12]. This error reduction is due to reduction in variance in the learned models [11]. Ensemble learning reduces errors most when the errors of the individual models are not highly correlated [1, 20]. Recent work [34] has shown that XAI methods for deep learners trained under slightly different circumstances produce explanations that are not highly correlated. The variability occurs due to the initial random parameter selection or the random order of training examples. This suggests that we can reduce the error of an explanation by combining explanations from an ensemble of networks.

Reiger [26] proposed combining different XAI methods such as Grad-CAM and LRP since each method has their own strengths and weakness and found this increased the stability of the XAI output. This does not use an ensemble of learners and cannot increase the accuracy of the classifications.

In learning from tabular data, ensembles have been shown to improve accuracy at the expense of interpretability. In contrast, our goal with image data is to both improve classification accuracy and the explanation.

3. METHODS

3.1. Generating Ensemble Explanations

For this paper, we use two common image classifiers: VGG16 [31] and ResNet [15]. We consider three methods of generating a diverse ensemble of classifiers.

1. **Different Random Weights**. We start with ***N*** identical base networks and then initialize each of the N classification heads with different random weights and present the same training data to each network. The idea here is that based on the initial conditions, the network will find a slightly different solution [3]. We evaluate whether on average the ensemble produces a better explanation than the members of the ensemble.
2. **Leave Out One Bucket.** We divide the training data into ***N*** buckets and train the ***N*** identical architectures on ***N-1*** buckets [14]. We evaluate whether the ensemble produces a better explanation than a single network trained on all the data.
3. **Bootstrap Aggregation.** We use bagging [4] which creates ***N*** training sets by sampling with replacement from the original training data. This leaves out some of the original training data and places additional weight on other examples by replicating them. We

evaluate whether the ensemble produces a better explanation than network trained on all the data. The motivation of the latter two methods is that slightly different training data will result in slightly different solutions that may be averaged. Even if each of the individual networks is less accurate than training on all the data, the consensus on the ensemble often exceeds the classification accuracy of a single network on all data. We anticipate this will hold with the accuracy of the explanation as well.

Once the models are trained, we generate an ensemble explanation by averaging the relevance score for each pixel (i, j) in the input image as shown in Eq. 1.

$$score_{ens}(i,j) = \frac{1}{n}\sum_{k=1}^{n} score_k(i,j) \tag{1}$$

3.2. Metrics

We consider three measures of explanation accuracy: Intersection over Union (IoU), correlation, and the center of mass distance. In all cases, the ground truth region is collected from human annotators. It is worth stressing that this region information is used only in evaluation, not in training.

1. **Intersection over Union** For IoU, we binarize the generated explanation by normalizing it between 0 and 1 (or [-1,1] for some XAI algorithms) and setting a threshold to find regions of interest. The IoU score is obtained by computing the intersection between the explanation and ground truth mask and then dividing it by their union. We use a default threshold of 0.3 in this work. Figure 3 illustrates how we compute the IoU score for a single image. The IoU metric allows us to compare how well an XAI algorithm identifies region of interest found by human annotators. A higher value shows more agreement between the algorithm and the annotator.

2. **Correlation** To quantify the similarities between explanation maps and ground truth masks, we consider the two as jointly distributed random variables and use the Pearson correlation between them. This metric is obtained by down sampling the masks to a lower resolution (e.g., 14×14) to reduce noise errors. Then we flatten the 2D masks into a 1D vectors and compute the correlation as:

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

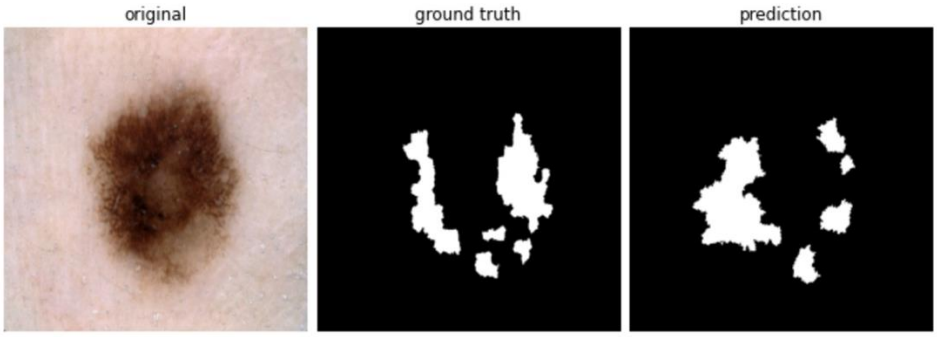


Figure 3. Left: Example of an image used to evaluate ensembles of explanations Middle: Ground-truth masks for an image in the ISIC-2018 melanoma dataset. Right: Explanation generated by averaging explanations generated by an ensemble of models.

3. **Center of Mass Distance** We compute the center of mass of a saliency or heatmap identified by the XAI algorithm. The center of mass ***R*** is computed as a function of each weight ***w*** at point ***r***.

$$R = \frac{1}{W}\sum_{i=1}^{n} w_i r_i \quad \text{where} \quad W = \sum_{i=1}^{n} w_i \tag{3}$$

Figure 3 illustrates the center of mass identified by ten networks for the wingbar of a bird. The black arrows indicate the center of mass of the heatmap of individual networks, and the red arrow indicates the center of mass of the ensemble explanation.

We compute the Euclidean distance from the center of mass to the key point identified by human annotator. The distance between the center of mass and a key point is useful for two reasons. First, it is quicker to collect key points vs. regions, i.e., pointing to a bird's bill vs. tracing it. Second, many explanations in medical journals or bird watching guides use arrows instead of heatmaps to identify features, and the center of mass can be used as the endpoint of the arrowhead.

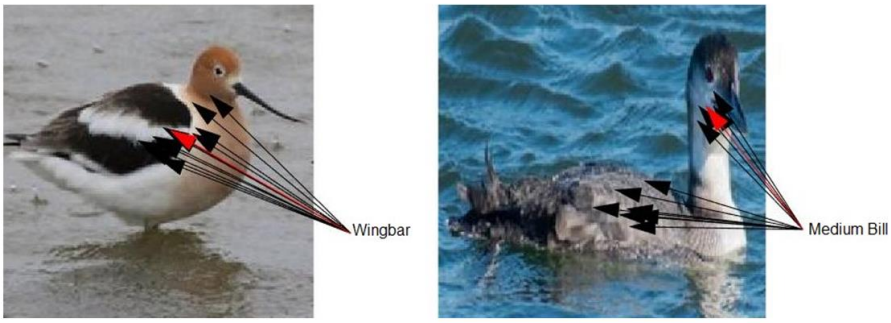


Figure 4. Use of key points that show center of mass (CoM) for heatmaps. The red arrowhead points to the CoM of the average heatmap while black arrows point to the CoM by each model in the ensemble.

3.3. XAI Methods

We evaluated the averaging approach on multiple XAI methods to show the generality of the approach. The methods that we explored are surveyed in [6] and include, GradCAM and Guided GradCAM [28], LIME [25], Input gradient [29], Gradient SHAP [21], Integrated Gradients [33], and Saliency [30].

4. DATASETS

We train and evaluate our averaging algorithm on two datasets. Note that a dataset for evaluation must include ground truth labels and ground truth regions or key points. The regions or key points are used in evaluation, but not in training.

ISIC-2018 ISIC-2018[8] is a melanoma detection dataset consisting of skin lesions obtained from a variety of anatomic sites and patients spread out across multiple different institutions. It consists of 2594 images along with 5 segmentation masks per image to identify the location of attributes of the region such as streaks and milia-like cysts. We only use the segmentation masks in evaluation. We train networks to recognize the presence or absence of features such as "milia-like cysts" and evaluate whether XAI algorithms find the region identified by the segmentation masks.

HiRes Birds HiRes Birds is a new dataset we introduce of 14,380 images of birds divided into 66 species. Additionally, we have collected data on various attributes of each bird, such as its bill length, wing pattern and location of the bill. We continue to collect additional feature and location information for this dataset. In this paper, we use this dataset to learn to identify attributes of the bird, such as whether it has a striped wing and evaluate whether the XAI algorithm focuses on the wing when making this classification.

5. RESULTS FOR ENSEMBLE LEARNING

In this section, we first present data that shows averaging explanations from an ensemble of learners improves the explanation for a variety of XAI algorithms using initial random weights to create the ensemble. Next, we show results that vary the conditions under which the ensemble is learned to go deeper into the conditions under which the approach is effective. To assess the performance of our proposed method, we define two quantities: Individual Average (Ind-Avg) and Ensemble Average (Ens-Avg). Individual Average is the average metric on the evaluation set of each of the individual networks in the ensemble in the case that the ensemble is trained with random weights, or the individual network trained on all the training data in the case of bagging and leave-out-one-fold. Ensemble Average calculates an average heatmap and we report the target metric between the average heatmap and the ground truth.

5.1. HiRes Birds Results

We consider two different classification tasks with the HiRes Birds dataset. In both cases, we create ensembles starting with different random weights.

First, we train on a multi-class problem of identifying the bill length of the birds. For this task, there are 3 classes: Large, Medium and Small. We train on 4763 examples from the dataset using 530 as a validation set and 2322 as an evaluation set. We used Hive Data (a crowd-sourcing platform for data labelling) to collect both the bill length data and the bill location data. Our dataset includes a key point for each bill.

Table 1 shows the distance between the center of mass of the region found by 5 XAI algorithms using VGG16 as the deep learning classifier. Table 2 shows the data using ResNet as the deep learning classifier. In these tables, statistically significant results using a paired t-test are shown in bold with p-value < 0.01 indicated by ** and p-value < 0.0001 indicated by ****. The results show that for five commonly used XAI algorithms and two commonly used deep learning architectures, our ensemble method results in better identification of the center of mass that can be compared to a key point.

In our second use of HiRes Birds, we train on a multi-class problem of distinguishing the wing pattern of birds. For this task, there are 6 classes. We train on 12221 examples from the dataset using 2156 as a validation set. We have collected wing patterns for our data, but not the wing location. Therefore, we evaluate on the location of the wing using the bird data from the PartImageNet which contains wing locations but not patterns. The evaluation is based on the ground truth locations of wings in 594 bird images from the PartImageNet dataset.

Table 1. Ensembling improves explanation center of mass distance on the beak size identification task with VGG16 [30]. **Ind-Avg** refers to average performance of individual models evaluated separately. **Ens-Avg** refers to performance of the ensemble model. Lower is better. Best results are in **bold**

Method	Ind-Avg	Ens-Avg
GradCAM	0.395	**0.328******
Input gradient	0.331	**0.320******
Gradient SHAP	0.346	**0.334******
Integrated Gradients	0.346	**0.334******
Saliency	0.328	**0.322****

Table 2. Ensembling improves explanation center of mass distance on the beak size identification task with ResNet18.

Method	Ind-Avg	Ens-Avg
GradCAM	0.236	**0.216******
Input gradient	0.315	**0.300******
Gradient SHAP	0.320	**0.309******
Integrated Gradients	0.330	**0.313******
Saliency	0.313	**0.309****

Table 3 shows the correlation between the importance of pixels identified by XAI algorithms and the wing region in PartImageNet using VGG16. As before, averaging over an ensemble improves the XAI algorithms we tested. In addition, we computed the center of the PartImageNet wing regions and compared to the center of mass of the regions found by various XAI algorithms. The results shown in Table 4 indicate improvement for this metric as well.

Table 3. Ensembling improves explanation correlation on the wing pattern identification task with ResNet18.

Method	Ind-Avg	Ens-Avg
GradCAM	0.207	**0.256******
Input gradient	0.263	**0.324******
Gradient SHAP	0.263	**0.315******
Integrated Gradients	0.265	**0.312******
Saliency	0.218	**0.337******

Table 4. Ensembling improves explanation center of mass distance on the wing pattern identification task with ResNet18.

Method	Ind-Avg.	Ens-Avg.
GradCAM [26]	0.157	**0.142******
Input gradient [27]	0.153	**0.145******
Gradient SHAP [19]	0.152	**0.146******
Integrated Gradients [31]	0.152	**0.146******
Saliency [28]	0.158	**0.147******

5.2. ISIC-2018 Results

We also tested our method on the ISIC2018 dataset for melanoma lesion detection. For this task, each training image is paired with 5 dermatologist annotated masks that identify important attributes for melanoma detection. We train a five different multi-label binary classifier starting with random weights to output a binary variable for each of the 5 attributes of an image. We generate the ground-truth binary labels for an image by checking whether the associated ground-truth mask is non-zero or not. Once trained, we run different XAI algorithms to generate regions for each attribute that we compare to the segmentation mask in the evaluating data. We evaluate the quality of the generated explanations of individual models against our proposed method which averages the explanation across the ensemble. Results are reported in Table 5. We observe that for GradCAM and LIME we see a noticeable and statistically significant increase in explanation quality with ensemble explanations. Integrated-gradients and Saliency show small but nonetheless statistically significant improvements. We believe this is because Integrated-gradients and Saliency generate pixel level explanations whereas GradCAM and LIME generate region-based explanations leading to better IoU scores when comparing regions to segmentation masks.

Table 5. IoU between the ground truth and explanations generated from XAI methods on ISIC 2018 dataset for the task of lesion identification.

Method	Ind-Avg.	Ens-Avg.
GradCAM	0.17	**0.19******
Integrated Gradients	0.04	**0.05******
Saliency	0.03	**0.04******
LIME	0.11	**0.13******

Our trained classifiers achieve a mean Area under the ROC curve (AUC) score of 0.82 on the validation set. Ensembling the models leads to an improved AUC score of 0.86, showing that in addition to increasing explanation accuracy, classification accuracy is also increased. The increase in accuracy is statistically significant with a p-value of 0.03.

5.3. Separating the Effect on Intersection and Union

Here, we investigate the effect of averaging explanations on the intersection and union metrics independently using the same training procedure described in section 5.2. A greater intersection means that the XAI algorithm finds more of the area of interest identified by an annotator. A smaller union indicates that fewer regions are found outside the relevant region. Both intersection and union are measured in pixels.

Table 6 shows that the average explanation increases the intersection and decreases the union for LIME, integrated gradients, and saliency. This indicates that averaging finds more relevant regions and fewer irrelevant regions. For GradCAM, both the intersection and union increase but the increase in the intersection is more significant, leading to an overall improvement in the IoU score.

Table 6. Intersection (I) and Union (U) metrics for melanoma dataset. Ind-Avg is the average of the metric of individual models. Ens-Avg is the metric for our proposed technique. The unit of measurement for both intersection and union is pixels.

Method	Ind-Avg-I	Ens-Avg-I	Ind-Avg-U	Ens-Avg-U
GradCAM	1344.82	1668.59	7968.59	8602.46
LIME	904.44	999.31	8262.70	7897.76
Integrated Gradients	95.20	118.46	3470.67	3463.11
Saliency	150.01	174.93	3574.14	3556.70

5.4. Other Approaches to Creating an Ensemble

We also tried training an ensemble on the HiRes Birds beaks dataset using the Leave-Out-One method of creating an ensemble. In this setting, we divide the training data into K = 10 folds and then train each model in an ensemble of size 10 with a unique combination of 9 folds. In a similar experiment, we trained another ensemble of size 10 using a bagging algorithm where the training set for each model is generated by sampling from the original training set with replacement. The results for both experiments are reported in Table 7. We observe that both K-fold and bagging lead to improved explanation accuracy when compared to explanations generated by individual models trained on the entire training data.

Table 7. Center of mass distance between explanations and ground-truth annotations for an ensemble trained using the K-fold and Bagging algorithms. Ind-Avg is the CoM distance of a single model trained on the entire training data. Lower is better.

Method	Ind-Avg	K-fold	Bagging
Guided GradCAM [26]	0.169	**0.153****	0.157
GradCAM[26]	0.297	**0.273****	0.277
Integrated Gradients [31]	0.278	**0.217****	0.219
Saliency [28]	0.280	**0.222****	0.225
Input gradient [27]	0.293	**0.241****	0.243
Gradient SHAP[19]	0.278	**0.218****	0.219

The results show that other ways of creating diverse models also works with our ensemble averaging method.

5.5. Varying the Size of the Ensemble

Next, we investigated the effect of ensemble size on the quality of the averaged explanation using the melanoma dataset. Fig. 5 plots the IoU score of the averaged explanation and the average IoU score of individual models against the size of the ensemble using the random weights method. We observe that IoU scores tend to increase with ensemble size and plateau after a certain threshold.

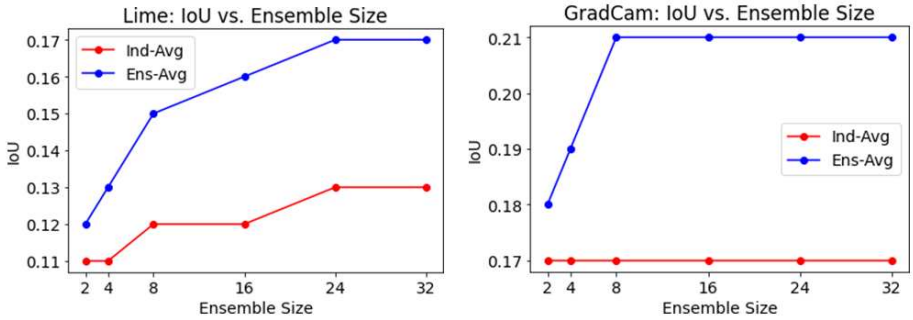


Figure 5. IoU between ground truth and generated explanations on ISIC-2018 dataset as a function of ensemble size. The blue curve shows the IoU score from our proposed technique which averages the explanations across the ensemble. The red curve shows the average of the individual IoU scores in the ensemble.

6. HUMAN EVALUATION

In this section, we report on three studies with human experts. The goals of these studies are:

1. To determine whether experts can detect the difference in the quality of explanations produced by ensembles when compared to individual models
2. To determine the format of explanations preferred by experts
3. To determine what type of explanation help novices learn the fastest

These studies were approved by UCSD's IRB.

6.1. Averaging Explanations

In this experiment, we show that people can notice the difference in the quality of explanations produced by ensembles when compared to individual models. In our computational experiments, the ensemble results in greater identification of regions of interest and less identification of irrelevant regions. The stimuli for the experiment consisted of images annotated by LIME or averaging of an ensemble of eleven LIME classifiers starting with different random weights. The stimuli were generated using the version of imageLIME in MATLAB. Figure 6 shows examples of the stimuli used in the experiment.

In Figure 6, the top bird is a common goldeneye. Its distinguishing characteristics (called field marks by birders) include a gold-colored eye, and it is distinguished from the similar Barrow's goldeneye by having a round vs. kidney-shaped patch on the cheek and striped vs. checkered wings. The averaged annotation on the right picks up both distinguishing characteristics, showing that averaging finds more relevant features. The middle bird is the Barrow's Goldeneye and again the ensemble focuses on both the wing pattern and the cheek patch. The lower bird is a western grebe. It is distinguished from the similar Clark's grebe by having a yellow vs. orange bill and having the black on the head extend below the eye. The average annotation on the right picks up both and furthermore does not emphasize a patch on the back that is not relevant to the classifications.

Participants for this study were expert bird watchers who were recruited from mailing lists that report rare bird sightings in Southern California. We recruited 28 participants for a LIME vs. averaged LIME study: one participant self-excluded due to a lack of familiarity with the bird species included in the studies and did not complete the study, and one was excluded due to being

under the age of 18, which may point to less real-world bird watching experience. In total, 26 participants were included in the analyses. Participants in the study had a median of 15 years of bird-watching experience.

Prior to beginning the study, participants were shown an example of LIME used on an image of a dog to identify the features most important to classifying its breed. This example was intended to familiarize participants with how to interpret the colors on a LIME-generated heatmap. Each study contained 24 unique bird images, each of which was shown once to each subject with LIME-generated annotations and another time with averaged LIME-generated annotations. This results in 48 trials evenly split between the base LIME method and the averaged LIME method.

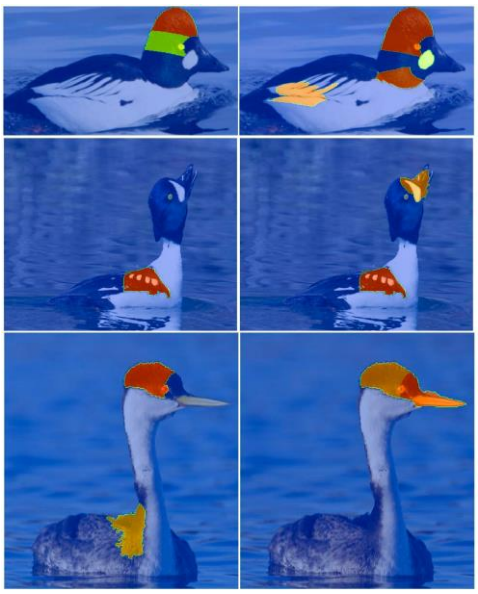


Figure 6. The figures on the left were annotated by LIME on a single VGG16 network. The annotations of the figures on the right are the average of 11 VGG16 networks.

In addition, these 24 unique bird images consisted of 10 different bird species, and these ten bird species were selected to include five pairs of similar-looking birds:

- Black-headed grosbeak and blue grosbeak.
- Clark's grebe and Western grebe.
- Eastern towhee and spotted towhee.
- Indigo bunting and lazuli bunting.
- The Barrow's goldeneye and common goldeneye.

Each trial displayed heatmap-annotated bird images alongside unannotated images, a bird species classification task, and questions about annotation preferences. A screen capture of the study interface is shown in Figure 7. Participants were asked to classify the bird species in the image by selecting 1 of 10 radio buttons corresponding to the ten unique bird species. In each study, participants were asked to provide their opinions on the novel highlighting method by answering two questions: "This highlighting emphasizes the areas of the bird that I think are important for identification" (Question 1), and "I would recommend using this highlighting to identify this bird" (Question 2). Participants indicated their responses using a 7-point Likert scale ranging from "Strongly Agree" (a value of 7) to "Strongly Disagree" (a value of 1). The midpoint (a value of 4) indicates a "Neutral" sentiment.

Figure 7. An example screen capture from the study.

For each of the two studies, we compared the median preference ratings and classification accuracy for standard LIME trials to averaged or contrasting LIME trials. We did not use classification accuracy as a premise for excluding trials or participants from analyses as classification accuracy may not correlate well to a person's ability to advise on whether an annotation is useful for classification purposes. For example, a seasoned bird watcher may recall that the color around the eye of a western or Clark's grebe is important for differentiating these two species; however, the bird watcher may not recall whether it is the western or Clark's grebe that has darker coloration around the eye. Thus, even though this bird watcher may perform below-chance at species classification, they are still able to point out which areas of a bird are integral to distinguishing between similar-looking species. Retaining erroneous trials did not result in a significant difference in the distribution of median ratings compared to excluding erroneous trials for either study (uncorrected p values > 0.7 for both studies). This is likely due to the high classification accuracy across all participants. Thus, all participants and trials were used for the subsequent analyses. The p values and median ratings for the two questions regarding preferences can be found in Figure 5. All reported p values are Bonferroni-corrected for 3 pairwise Wilcoxon signed-rank tests in each study.

Annotations averaged over an ensemble were significantly preferred to standard LIME-generated annotations for both Q1 and Q2 ($p < 0.001$). The median rating of Q1 for LIME annotations was 4.0 ("Neutral") while the median rating for averaged LIME annotations was 5.5 ("Slightly Agree"/"Agree"). The median rating of Question 2 for the LIME annotations was 3.0 ("Slightly Disagree") while the median rating for the averaged LIME annotations was 5.0 ("Slightly Agree"). With a mean of 90.1% and 91.3% for LIME and averaged LIME, respectively, the type of annotation did not make a difference with respect to bird species classification accuracy ($p >$ 0.99). Subjects in general thought the ensembles led to improved highlighting areas that are important for identification and would recommend using that over the regions identified by a single model.

6.2. A Broad Comparison of Annotation Types

We recruited 21 expert bird watchers via mailing lists that report rare bird sightings in Southern California to obtain their feedback on the utility of 5 types of image annotations for bird classification: bounding boxes, arrows, arrows with labels, heatmaps, and verbal descriptions. The heatmaps in this study were produced by using VGG16 as the classification algorithm and Grad-CAM as the XAI algorithm. Before beginning the study, participants were shown an example of Grad-CAM used on an image of a dog to identify the features most important to classifying its breed. This example was intended to familiarize participants with how to interpret the heatmap (i.e. the red to blue color gradient overlaying the image signifies the transition between the more and less important areas in classifying the dog's breed).

The study interface is shown in Figure 8: each trial in the study shows an image of a bird annotated with one type of aforementioned annotation on the left-hand side of the screen. Participants were able to toggle between the annotated and original, unannotated image by clicking the bird image on the screen. Participants saw 10 trials of each type of annotation for a total of 50 trials, each trial showing a unique bird image annotated with 1 of the 5 annotation types. On the right-hand side of the screen, participants were asked to answer two questions in each trial to gauge preferences for annotation types: "This explanation emphasizes the areas of the bird that I think are important for identification" (Q1) and "I would recommend using this explanation to help identify this bird" (Q2). These two questions were intended to gauge whether the emphasis was on the correct areas of the bird (Q1) and whether the explanation is useful not necessarily to only the expert but for the general end-user (Q2). A third question, "I am confident in my answers to the above questions" (Q3), was asked in each trial to potentially exclude uncertain participants or individual trials. Participants were asked to answer these questions on a 7-point Likert scale ranging from "Strongly Disagree" (a value of 1) to "Strongly Agree" (a value of 7) with "Neutral" (a value of 4) at the midpoint. An example of the 5 types of annotations used are in Figure 9.

Figure 8. An example of a heatmap trial in our first study, which compares preferences for 5 different annotation methods among birders.

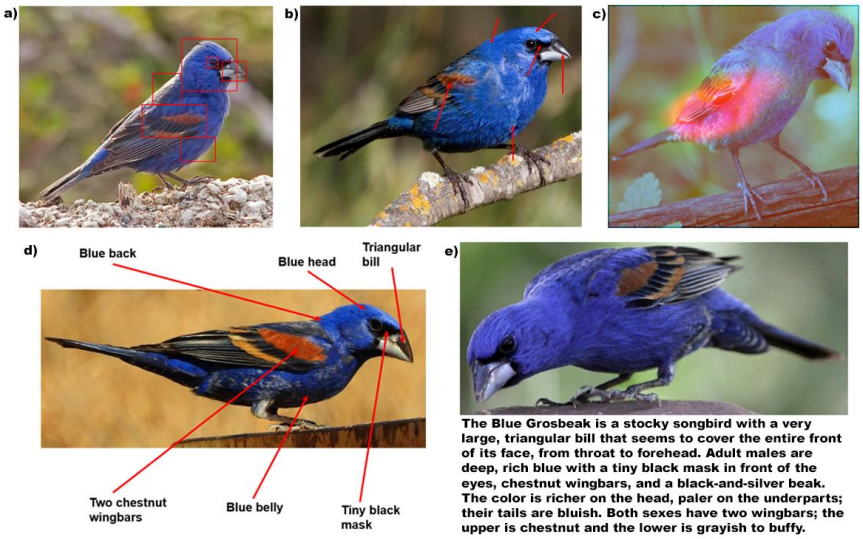


Figure 9. Examples of different kinds of annotations applied to the blue grosbeak: a) bounding boxes, b) arrows, c) heatmaps, d) labeled arrows, and e) a verbal description.

We calculated the median ratings for Q1 and Q2 for each type of annotation for each participant. No participants or trials were excluded from these analyses due to high response confidence (Q3) across all participants (median = 7.0 for all annotation types). We conducted a pairwise Mann-Whitney U test for all combinations of annotation types, the results of which can be found in Figure 10 and the full results in Table 8. All p values are Bonferroni-corrected for 20 pairwise comparisons.

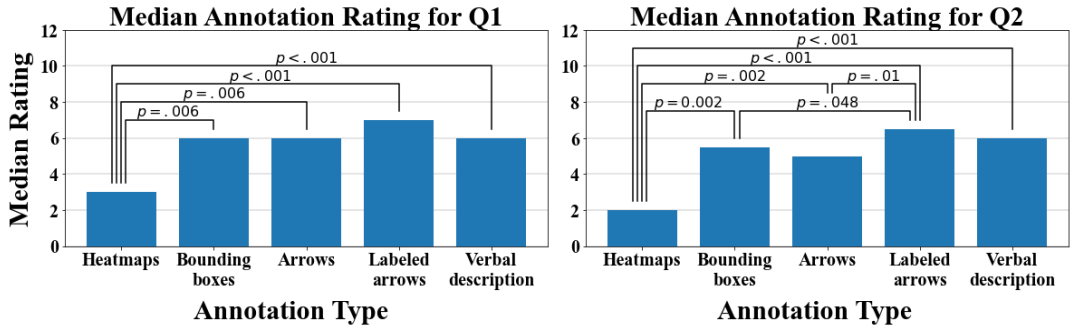


Figure 10. Experts' median preference ratings for Q1 and Q2 where $p < .05$, $p < .01$, and $p < .001$ are signified by *, **, and ***, respectively.

Table 8. Bonferroni-corrected *p* values for Q1 and Q2 for pairs of explanations.

Explanation Comparison	Question	
	Q1	Q2
Heatmaps vs. bounding boxes	0.006	0.0019
Heatmaps vs. arrows	0.006	0.002
Heatmaps vs. labeled arrows	<0.001	<0.001
Heatmaps vs. verbal description	<0.001	<0.001
Bounding boxes vs. arrows	0.99	0.99
Bounding boxes vs. labeled arrows	0.27	0.048
Bounding boxes vs. verbal description	0.99	0.99
Arrows vs. labeled arrows	0.053	0.012
Arrows vs. verbal description	>0.99	>0.99
Labeled arrows vs. verbal description	0.87	0.89

Labeled arrows were the most preferred type of annotation with median ratings of 7.0 for Q1 and 6.5 for Q2. In contrast, heatmaps were the least preferred type of annotation and the only annotation type to garner overall negative sentiment with a rating of 3.0 (“Slightly Disagree”) for Q1 and 2.0 (“Disagree”) for Q2. All other types of annotations saw a minimum rating of 5.0 (“Agree”) for either Q1 or Q2. Consequently, there was a significant difference in median preference ratings between heatmaps and all other annotation types for both Q1 and Q2, with the most drastic difference being between labeled arrows and heatmaps. There were fewer significant differences between bounding boxes, arrows, arrows with labels, and verbal descriptions than there were between heatmaps and four previously mentioned annotations.

6.3. Explanations to Assist Novice Learning

A final user study looks at the effect of different types of image annotations in helping novice birders become proficient at classifying similar-looking bird species. 336 students were sourced from UCSD’s undergraduate subject pool of students enrolled in psychology, linguistics, or cognitive science courses and screened for prior experience with birdwatching or identification. The subjects were initially evenly counterbalanced between the four annotation types in which we were interested (i.e. heatmaps, bounding boxes, arrows with labels, and verbal descriptions) and the control group (no annotations). As participant recruitment progressed, we transitioned to collecting data specifically on the control, labeled arrows, and heatmap conditions in order to better compare what we have seen to be the most and least preferred annotation type by experts.

The objective of this study was to determine which types of image annotations given as feedback for a classification task help minimize the number of trials a naive subject takes until they are proficient at distinguishing a pair of similar-looking birds. Proficiency is defined as the correct classification of 9 out of 10 birds in a running window of 10 trials. Due to screening out prospective participants with prior birdwatching or identification experience, all subjects must begin the study with a trial-and-error approach to correctly classifying the bird on the screen.

Each trial is a binary classification task with one option being the name of the (correct) bird shown on the screen and the alternative being the name of an (incorrect) lookalike bird. Importantly, the birds used in this were given pseudonyms (e.g. western towhee) as some bird names contain clues as to identifying the bird (e.g. spotted towhee). After submitting their classification, all subjects are given feedback on whether their classification was correct or not and shown the name of the correct bird. Additionally, the experimental groups are shown the image of the bird they classified along with annotations that emphasize areas of the bird important to its correct classification. Participants in the control group do not receive any annotations on the image. Examples of the feedback given to the labeled arrow experimental group and control group is shown in Figure 11.

Participants repeat this cycle of classification and feedback with (or without) annotations until they are able to correctly classify 9 out of 10 birds in a running window of 10 trials. There were a total of 20 unique bird images for each pair of lookalike birds; thus, if a participant requires more than 20 trials to reach proficiency, these bird images are reshuffled and reused. After reaching classification proficiency, the participant must take a brief one minute break before moving on to the next pair. There were a total of three bird pairs used in this study.

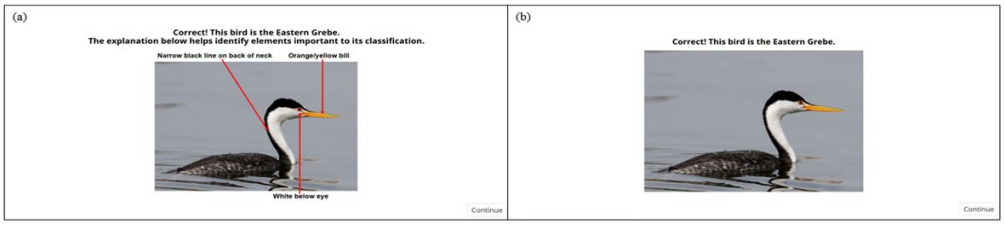


Figure 11. (a) An example of feedback a subject in the experimental group given labeled arrows would receive after a trial. (b) An example of the absence of feedback subjects in the control group receive after a trial.

Accounting for the possibility that some bird pairs may be more difficult to distinguish than others, we calculated the standard deviation and median separately for each bird pair and used these values to establish an exclusion criteria: participants requiring more than 2 standard deviations above the median number of trials until proficiency for a given bird pair were excluded from further analysis related to that bird pair. This measure was intended to filter out inattentive participants.

For all three bird pairs, there was no significant difference between heatmap annotations and having no annotations ($p > 0.91$ for all bird pairs) as feedback. Labeled arrow feedback consistently led to a lower number of trials until proficiency compared to heatmaps ($p < 0.05$ for goldeneyes, $p < 0.01$ for towhees, and $p < 0.001$ for grebes) and no annotations ($p < 0.001$ for all bird pairs); moreover, labeled arrow annotations consistently achieved the smallest or tied for smallest number of trials until proficiency for all bird pairs. The median number of trials for bounding box and verbal description annotations tended to fluctuate between bird pairs, likely due to a lower sample size for these conditions. The median number of trials taken until classification proficiency as well as significant differences between experimental groups are shown in Figure 12 and a full table of p values is given in Table 9.

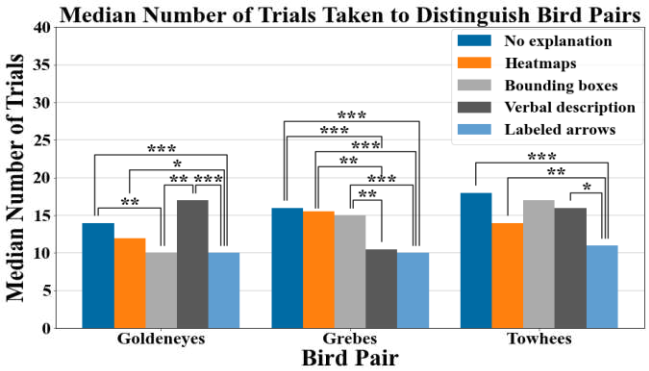


Figure 12. Median number of trials until bird classification proficiency. Asterisks *, **, and *** signify $p <$ .05, $p <$ .01, and $p <$ .001, respectively.

Table 9. Bonferroni-corrected *p* values for bird pairs by pairs of explanations.

Explanation Comparison	Bird Pair		
	Goldeneyes	Grebes	Towhees
No explanation vs. heatmaps	0.91	>0.99	>0.99
No explanation vs. bounding boxes	0.002	>0.99	>0.99
No explanation vs. labeled arrows	<0.001	<0.001	<0.001
No explanation vs. verbal description	>0.99	<0.001	>0.99
Heatmaps vs. bounding boxes	0.25	>0.99	>0.99
Heatmaps vs. labeled arrows	0.03	<0.001	0.0098
Heatmaps vs. verbal description	0.11	0.002	>0.99
Bounding boxes vs. labeled arrows	>0.99	<0.001	0.08
Bounding boxes vs. verbal description	0.002	0.005	>0.99
Labeled arrows vs. verbal description	<0.001	>0.99	0.01

In summary, we found that labelled arrows that include descriptions of regions were preferred by experts and helped novices learn fastest. We also explored two extensions of XAI systems, contrastive explanations and ensembles of explanations and found experts preferred these extensions to the base algorithms. We explored contrastive explanations extensively in prior years. In the final year, we also explored averaging over multiple explanations.

The three experiments reported here demonstrate that people prefer explanations that include semantic labels on regions and can detect the difference between explanations created by ensembles to individual models. In the next section, we describe our method for labeling images.

7. MULTITASK LEARNING TO ADD EXPLANATORY LABELS TO IMAGES

We developed an approach to explaining image classification by annotating images with labeled arrows and describing regions of the image that are important to the classification. Figure 13 illustrates the network architecture we are using. We use multitask learning and simultaneously train the network on a class label and whether or not the image contains features such as “short bill," “long bill," “solid wing," wingbars," etc. Optionally, we can also add hierarchical information such as the family of the bird or whether or not a mole is cancerous in addition to the specific type of cancer such as “subcutaneous melanoma.”

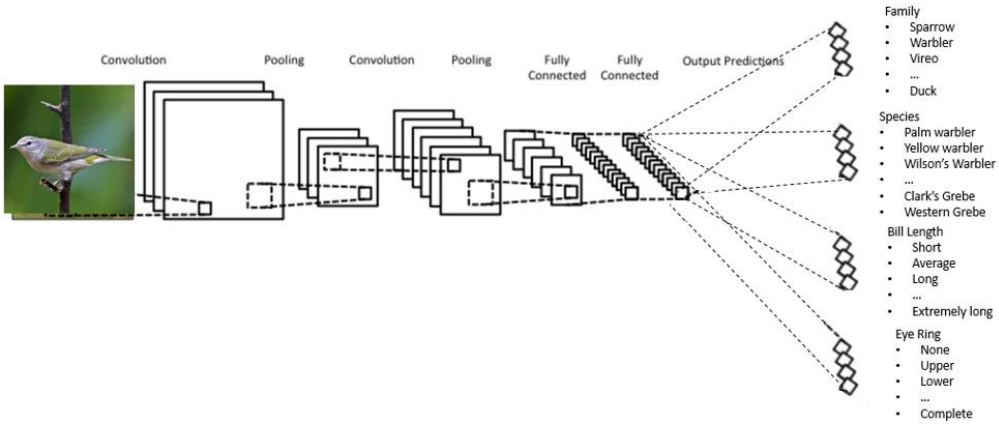


Figure 13. Network architecture for multitask learning on the bird species and physical attributes.

We rely on saliency maps generated by any XAI algorithm to indicate the regions important to explaining why descriptive features such as “wingbars" are present in the image. However, rather than superimposing a heatmap on the image, we find the center of mass of the saliency

map and use that as the end point of a labeled arrow. Note that the multitask method can easily be combined with the ensemble learning described earlier. To use multitask learning, images need to be annotated not only with class info but also descriptions of parts as we have done with the HiResBirds where we have collected info on various attributes, such as the length of the bill and the pattern on the wing and polygons outlining the bill and the wing. Note that the polygons are not used in training, only in evaluation. In some cases, e.g., the eye, we collect a keypoint instead of a polygon.

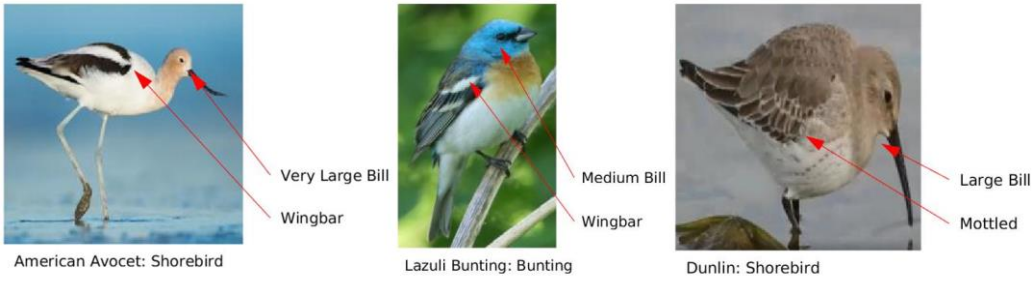


Figure 14. Automatically labeled image explaining bird classification.

Figure 14 shows an example annotation of a bird using our system with VGG16 as the network architecture and GradCAM as the XAI algorithm. We have obtained similar results with other XAI algorithms and other network architectures. This method generates explanations by labeling images in a manner similar to the explanation produced by experts.

8. CONCLUSION

XAI algorithms were developed to increase trust in deep learning algorithms for tasks such as image classification. However, XAI algorithms themselves need to be trustworthy. It has been shown that differences in training and initial conditions can produce different explanations and that the explanations of current XAI systems fail to identify all regions of importance used by human experts.

In this paper, inspired by the success of ensembles to increase classification accuracy, we proposed using ensembles to improve the explanation accuracy of saliency-based XAI algorithms. We show, through empirical results, that ensembles can improve the accuracy of explanations when measured using metrics such as IoU, correlation, and center of mass distance. Furthermore, we showed that explanations produced by ensembles are preferred by people over explanations produced by a single network. By looking for areas of consensus across multiple networks, ensembles reduce the irrelevant areas and increase the relevant areas in explanation.

Furthermore, by using multitask learning and computing the center of mass we can produce explanations with semantically labelled arrows in a format similar to that used by experts.

ACKNOWLEDGEMENTS

This work was supported with funding from the DARPA Explainable AI Program under a contract from NRL and from NSF grant 2026809. We would like to thank Dave Gunning for stimulating XAI research. Discussions with David Aha, Albert Hsiao, and Justin Huynh were helpful in developing the ideas in this paper.

REFERENCES

[1] K. M. Ali and M. J. Pazzani, "Error reduction through learning multiple descriptions," *Machine learning*, vol. 24, no. 3, pp. 173–202, 1996.

[2] N. Arun, N. Gaw, P. Singh, K. Chang, M. Aggarwal, B. Chen, K. Hoebel, S. Gupta, J. Patel, M. Gidwani *et al.*, "Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging," *Radiology: Artificial Intelligence*, vol. 3, no. 6, p. e200267, 2021.

[3] P. P. Brahma, D. Wu, and Y. She, "Why deep learning works: A manifold disentanglement perspective," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 10, pp. 1997–2008, 2016.

[4] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[5] Caruana, R. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[6] S. Chakraborty *et al.*, "Interpretability of deep learning models: A survey of results," in *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*. IEEE, 2017, pp. 1–6.

[7] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao *et al.*, "Interpretability of deep learning models: A survey of results," in *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*. IEEE, 2017, pp. 1–6.

[8] N. C. F. Codella *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC)," *CoRR*, vol. abs/1902.03368, 2019. [Online]. Available: http://arxiv.org/abs/1902.03368

[9] L. A. de Souza Jr, R. Mendel, S. Strasser, A. Ebigbo, A. Probst, H. Messmann, J. P. Papa, and C. Palm, "Convolutional neural networks for the evaluation of cancer in barrett's esophagus: Explainable ai to lighten up the black-box," *Computers in Biology and Medicine*, vol. 135, p. 104578, 2021.

[10] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.

[11] P. Domingos, "A unified bias-variance decomposition," in *Proceedings of 17th International Conference on Machine Learning*, 2000, pp. 231–238.

[12] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 241–258, 2020.

[13] Y. Freund, R. E. Schapire *et al.*, "Experiments with a new boosting algorithm," in *icml*, vol. 96. Citeseer, 1996, pp. 148–156.

[14] M. Gams, "New measurements highlight the importance of redundant knowledge," in *Proceedings of the 4th European Working Session on Learning (EWSL89)*, 1989, pp. 71–80.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2016, pp. 770–778.

[16] Jonas, J., Cursiefen, C., and Budde, W. Optic Neuropathy Resembling Normal-Pressure Glaucoma in a Teenager With Congenital Macrodiscs. *Arch Ophthalmol*.;11610:1384–1386. 1998

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[18] S. Lapuschkin, A. Binder, G. Montavon, K.-R. Müller, and W. Samek, "The lrp toolbox for artificial neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3938–3942, 2016.

[19] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[20] Y. Liu and X. Yao, "Ensemble learning via negative correlation," *Neural networks*, vol. 12, no. 10, pp. 1399–1404, 1999.

[21] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[22] S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," *CoRR*, vol. abs/1705.07874, 2017.

[23] Mar, V.J., Soyer, H., Button-Sloan, A., Fishburn, P., Gyorki, D.E., Hardy, M., Henderson, M. and Thompson, J.F., 2020. Diagnosis and management of cutaneous melanoma. *Australian journal of general practice*, *49*(11), pp.733-739.

[24] M. Pazzani, R. K. Severine Soltani, S. Qian, and A. Hsiao, "Expert-informed, user-centric explanations for machine learning," in *Proceedings of the AAAI Conference on Artificial Intelligence-2022*. IOS Press, 2022.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "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, pp. 1135–1144.

[26] L. Rieger and L. K. Hansen, "Aggregating explainability methods for neural networks stabilizes explanations," *arXiv preprint arXiv:1903.00519*, 2019.

[27] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[29] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[30] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[32] A. Singh, S. Sengupta, A. R. Mohammed, I. Faruq, V. Jayakumar, J. Zelek, V. Lakshminarayanan *et al.*, "What is the optimal attribution method for explainable ophthalmic disease classification?" in *International Workshop on Ophthalmic Medical Image Analysis*. Springer, 2020, pp. 21–31.

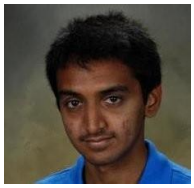
[33] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[34] M. Watson, B. A. S. Hasan, and N. Al Moubayed, "Agree to disagree: When deep learning models with identical architectures produce distinct explanations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 875–884.

[35] X. L. Weina Jin and G. Hamarneh, "Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements?" in *Proceedings of the AAAI Conference on Artificial Intelligence-2022*. IOS Press, 2022.

[36] https://www.canada.ca/en/environment-climate-change/services/migratory-game-bird-hunting/frequently-asked-questions/barrow-goldeneye-recognize-it.html (Accesses 11/25/2022).

AUTHORS

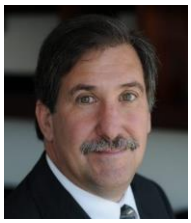


Aadil Ahamed received his M.S. in Computer Science, Specialization in Artificial Intelligence from University of California, San Diego and his B.S. in Computer Engineering from University of California, Irvine.

Kamran Alipour received his PhD in Computer Science from University of California, San Diego in 2022 and M.S. in Aerospace Engineering from Sharif University of Technology and a B.S. in Aerospace from K. N. Toosi University of Technology

Sateesh Kumar is a Master's student at the Computer Science and Engineering department of UC San Diego. He obtained his bachelor's in Computer Science from National University of Computer and Emerging Sciences, Pakistan.

Severine Soltani completed her B.S. in Cognitive Science with a focus on machine learning and neural computation at UC San Diego in 2020. She is currently a Bioinformatics and Systems Biology Ph.D. student at UC San Diego, examining the effects of environmental perturbations on physiological data via wearable devices.



Michael Pazzani received his Ph.D. in Computer Science from University of California, Los Angeles. He is director of the Artificial Intelligence Research for Health Center at the Information Sciences Institute in Marina Del Rey, CA, USA