# Interpreting Deep Learning: ISIC Melanoma

**Matteo Russo**
Princeton University
matteor@princeton.edu

**Dale Lee**
Princeton University
dalelee@princeton.edu

**Andrew Griffin**
Princeton University
andrewjg@princeton.edu

## Abstract

Can an algorithm save a life? In the field of dermatology, an early diagnosis can make the difference between life and death. Skin cancer is the most prevalent kind of cancer in the United States, with 1 in 5 Americans developing skin cancer by the age of 70 [3]. Melanoma is the most fatal type of skin cancer, but can often be cured if detected early enough, with the 5 year survival rate for early-stage melanoma being 99% [4]. This unique aspect of this particular kind of cancer makes it ideal for a computer vision application since potential lesions are easy to identify, image, and visually diagnose, unlike other kinds of internal cancers that require a more vigorous biopsy procedure. In this project, we establish baseline classification techniques on melanoma image classification, data exploration and visualization via dimensionality reduction, as well as exploring Convolutional Neural Network approach, providing interpretability explanations of the predictions through three methodologies: Activation Maps Backpropagation, LIME and SHAP values. We found the highest performance with the CNN approach. We also found that the interpretable predictions truly shed light on the model and onto the degree to which we should trust the Deep Learning prediction, both of which are crucial in the field of bio-medicine.

## 1 Introduction

This report will show a natural progression through various machine learning techniques in order to complete the important task of classification of skin lesion images as benign or malignant. There will also be an emphasis on interpretation of models, especially given the fact that the classification task is currently being done solely by human doctors and a potential switch from humans to algorithms will not be reliable unless backed up by interpretations that help explain what the algorithm is picking up from the image data. Even without the possibility of replacement of doctors by algorithms, interpreting these models can give doctors a very powerful tool, and complement the biology and anatomy that has been studied on the topic. The progression towards effective classification of images will begin with the baseline classification techniques of random forest and logistic regression. These techniques are effective at classification of many types of data including text data for sentiment analysis, data on students for educational outcomes[6], etc. These techniques are general and give us something to compare to the results of our more complicated models. We see that the general classification techniques are not great, yielding only about 77% accuracy on 224x224 pixel images. This motivates us to use other techniques to analyze the data and attempt to gain a greater understanding for the underlying structure and patterns.

This will bring us to unsupervised learning in the form of dimension reduction techniques and visualizations. By performing PCA and visualizing the corresponding clusters with respect to the first

two principle components, we see that there is a large amount of class overlap. This confirms our intuition that a linear classifier will not suffice for the data in question, and further enforces the hypothesis that we need a better classifier.

Finally, we will implement a more complicated classifier using a Convolutional Neural Network (CNN) in order to gain greater accuracy in classification. CNNs are great for image data as they take advantage of local patterns and spatial considerations that exist in images. The one disadvantage of using a complex nonlinear model such as CNN is the difficulty of interpretability. As previously stated, this interpretability is key in the medical field where insights must be backed up by reliable explanations. These interpretations will come in the form of visualizations of algorithms assigning importance to different pixels and objects in an image. The effectiveness of the CNNs compared to the baseline classifiers with the corresponding visual interpretations give us compelling evidence that machine learning techniques *can*, in fact, be extremely useful for medical applications such as melanoma detection.

We used the HAM10000 ISIC 2017 dataset from the International Skin Imaging Collaboration Melanoma Project, which is a public archive of dermoscopic images of skin lesions [2]. Because computation with RGB image data can be quite time complex, we used a reduced dataset of 1449 224x224px RGB malignant images and 1800 224x224 px RGB benign images via random sampling. We scaled the original images which were 500x1000px to 224x224px, using using reverse mapping from the destination to source image with the cv2 resample library. This method of image reduction has found success in many implementations [1].

## 2 Related Work

Neural Networks are widely used in image classification—in particular, convolutional neural networks have been found to reduce computation time while reducing the chance of overfitting and providing improved results on smaller datasets. Convolutional Neural Networks in the biomedical imaging sphere have gained widespread success in classifying and segmenting images [9]. A recent study published in Nature by Stanford researchers found that a CNN was able to outperform board-certified dermatologists when using a pre-trained GoogleNet Inception v3 CNN [5].

There is also literature on fine-tuning an already trained network for image classification versus creating one from scratch.[13] Because of issues with computational power and the need for large data sets, many researchers use existing networks trained on other images for classification. We believe melanoma images are sufficiently different from general image training data, therefore we choose to fully train rather than fine tune an existing network.

## 3 Methods

### 3.1 Baseline Classification Methods

We use two standard classifiers to benchmark a baseline accuracy for the ISIC Melanoma Classification Task:

- Random Forest with a max tree depth of 5 and number of estimator trees of 20.
- Logistic Regression with $\ell_2$ regularization.

We found the above mentioned hyper parameters through 10 fold Grid-Search Cross Validation. We searched over 5, 10, 20, 35, 50 number of trees in the random forest, with tree depths of 3 and 5. For logistic regression, we searched over l1 and l2 penalty space.

### 3.2 Dimensionality Reduction and Data Visualization

Due to the large number of features we test on, our image classifier could suffer from the "curse of dimensionality" and lead to overfitting of the training data due to the high number of irrelevant parameters [11]. We decided to test this possibility by performing PCA with two principal components, and visualizing the results to see what the principal components were in Figure 9. We did

not find a stark separation of clusters with PCA, however, so we decided to perform t-SNE projection to two dimensions, and found a sharper divide between malignant and benign dimensionally reduced lesions. We recognize the fact that t-SNE performs well with low-dimensional samples and that is the reason why we decided to first reduce the dataset through PCA and then apply t-SNE to the linearly reduced dataset. Even in this case, nonetheless, we acknowledge that this pipeline of dimension reduction has no theoretical ground but it gives similar performances in the visualization to the ones t-SNE provides. We could visualize t-SNE reduction in Fig.1, PCA reduction in Fig.9 and PCA followed by t-SNE in Fig. 11 and Fig. 12.

### 3.3 Convolutional Neural Network (CNN)

A Convolutional Neural Network is a machine learning algorithm meant to capture the importance of different aspects of images to their classification. It performs remarkably well with image data as it takes into account spatial consideration of matrices (pixels) by applying filters to subsets of the pixels, then pooling together the output of those filters in order to reduce dimensions and capture the most important features.

### 3.4 CNN Visualizations

We will see in Section 4 that convolutional neural networks perform much better at the classification of benign and malignant lesions in the ISIC dataset. However, because the neural network is very complex and highly nonlinear, it can be difficult to interpret. Specific to a medical dataset like ISIC, this interpretation is paramount. It is unlikely that professionals in the medical field will be comfortable with trusting the output of a model where nobody can understand exactly why classifications are being made. The interpretability of a complex model, in general, is as important if not more important than the model prediction itself. It is often the case that there exists a true dichotomy between the complexity of the model, and thus, the hardness of the patterns it is able to capture and its interpretability. As is the case with many machine learning techniques, visualizations can shed light on the most important aspects of the model. The visualizations that are used in this report are class activation maps as well as LIME and SHAP visualizations.

#### 3.4.1 Class Activation Maps

Class Activation maps allow us to see which areas in the image are relevant to the classification of a specific class. For example, CAMs are used in order to show the importance of faces and limbs in the classification of humans along or the importance of ears and noses in the classification of dogs and other animals.[14] As highlighted in Section 3.4, Neural Networks are constructed in convolutional layers created by different filtering of images by a kernel. Each convolutional layer extracts high-level features such as edges or patterns in the input image, and by analyzing CAMs at each layer we can get an appreciation for how each layer extracts a different feature. Some layers clearly extract the shape of the objects in the image, other layers clearly extract color, and yet others extract combinations of features that are not as easily recognizable. We will look at these CAMs with respect to properly classified samples as well as misclassified samples to see exactly how the CNN model assigns importance to areas in the picture for classification. In other words, the Activation Map Analysis we have carried out relies on a white-box knowledge of the network itself, meaning that we have perfect knowledge over all the components. As a consequence, albeit their strong visual interpretability, Activation Maps do not scale well to other models.

#### 3.4.2 LIME

Urging to abstract oneself from the perfect knowledge of the network, Ribeiro et al.'s LIME (Local Interpretable Model-agnostic Explanations [10]) performs a black-box locally interpretable model evaluation by perturbing, within a neighborhood, specific samples' feature values and measuring the resulting impact in the classification. Formally, let $g \in \mathcal{G}$ be an interpretable model from the set of interpretable models $\mathcal{G}$ and let $f$ be the original globally unknown model. Further, let $\Omega(g)$ be a measure of model $g$ complexity (such as Radamacher's Complexity) and let $\pi_{\boldsymbol{x}}$ be a definition of locality (e.g. neighborhood) for point $\boldsymbol{x}$. Last, let $\mathcal{L}(g, f, \pi_{\boldsymbol{x}})$ be the measure of mistrust we have on model $g$ in approximating $f$ within locality $\pi_{\boldsymbol{x}}$. Then, LIME solves the following minimization

problem:

$$\lambda(\boldsymbol{x}) = arg\min_{g \in \mathcal{G}} \mathcal{L}(g, f, \pi_{\boldsymbol{x}}) + \Omega(g) \tag{1}$$

The notion of locality $\pi_{\boldsymbol{x}}$ in $\mathcal{L}$ allows us not to worry about the overall model $f$ but to only consider a neighborhood approximation of it.

#### 3.4.3 SHAP

As much as LIME, with the aim of visual interpretability, rather than explaining the complex model $f$ in its entirety, Lundeberg and Lee in [8] propose to study the model $f$ itself when evaluated at a single data point, as per the following equation, where $g$ represents the additively binary simplified version of $f$, $\boldsymbol{\phi}$ the vector of feature coefficients for each of the covariate components and $\mathbb{I}$ is the indicator which is equal to 0 if the feature is not present and 1 if it is:

$$g = \boldsymbol{\phi}^T(\boldsymbol{x})\mathbb{I}(\boldsymbol{x}) \tag{2}$$

The only question that remains unanswered regards how to choose such that the following four notions of fairness are fulfilled for any sample $\boldsymbol{x}$:

1. *Perfect redistribution*: $g(\boldsymbol{x}) = f(\boldsymbol{x})$.
2. *Equality*: features contributing equally to the classification should have the same SHAP value.
3. *Neutrality*: features not contributing should have null SHAP value.
4. *Additivity*: $g(\boldsymbol{x} + \boldsymbol{y}) = g(\boldsymbol{x}) + g(\boldsymbol{y})$.

The following expression gives us the feature coefficients that would satisfy the four properties of above:

$$\phi_i = \sum_{\mathcal{S} \subseteq \mathcal{F} \setminus \{i\}} \frac{|\mathcal{S}|!(|\mathcal{F}| - |\mathcal{S}| - 1)!}{|\mathcal{F}|!}[f_{\mathcal{S} \cup \{i\}} - f_{\mathcal{S}}] \tag{3}$$

Hereby, $\mathcal{F}$ represents the set of all features, $i$ represents the ith component of feature map $\boldsymbol{\phi}$ and $\mathcal{S}$ any subset of $\mathcal{F} \setminus \{i\}$. In other words, SHAP calculates the importance of each feature (in our case, pixel), in any possible order so to avoid unfair order bias in features comparison. As we may observe from Eq.2, the problem of finding every possible subset $\mathcal{S} \subseteq \mathcal{F} \setminus \{i\}$ and retraining on each of them explodes combinatorially, and, thus, we need convex approximation techniques to gain computational efficiency. For a Deep Neural Network model $f$, as thoroughly explained in [8], SHAP is replaced by a technique named Deep SHAP. The latter links Shapley values computed in small local portions of the DNN to the ones for the whole network, via a recursive backpropagation of DeepLIFT multipliers ([12]).

## 4 Results

### 4.1 Exploratory Analysis: PCA and t-SNE Feature Reduction

When observing the PCA transformed dataset pictured in Figure 9, we found that the two directions of maximum variance appeared to be color and size—images that were more blue appeared to cluster higher in the plot, while larger lesion images tended towards the right side of the plot, and smaller lesions tended towards the left. However, PCA by itself did not provide a very clear separation between the malignant and benign lesions. Because the blue tinted images seem to correspond most with the cluster of benign images, we considered the fact that this could be an artifact of the data, however, we found that matplotlib was modifying the background color of the images, but the correctly colored images were being input to the classifiers.

We found improved separation between malignant and benign lesions through first applying PCA, then t-SNE to the dataset, as shown in Figure 11 and Figure 12 in the Appendix. PCA is only taking the principal components of the images, while t-SNE is perhaps a better application to nonlinear data. This was confirmed when we applied t-SNE by itself to the dataset, finding that this dimensionality reduction methodology offered the clearest separation between the two kinds of lesions.

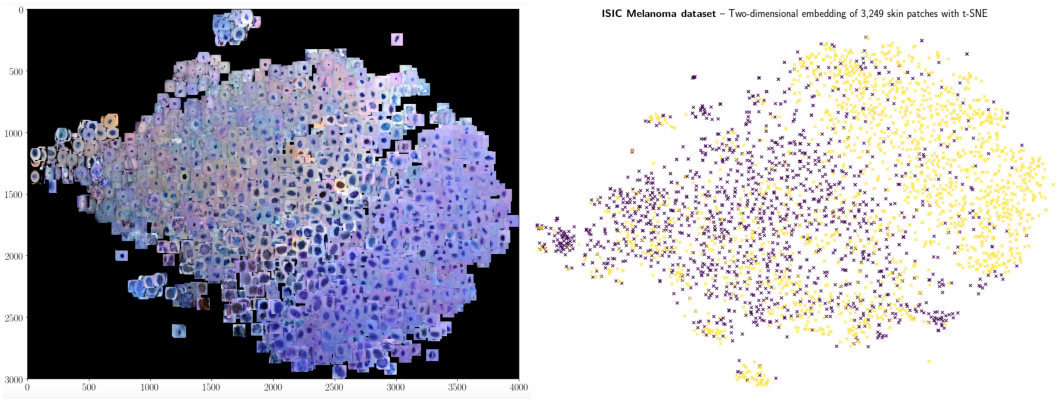



Figure 1: t-SNE of benign and malignant images. Yellow datapoints are benign, and black datapoints are malignant.

### 4.2 Baseline Classification Results

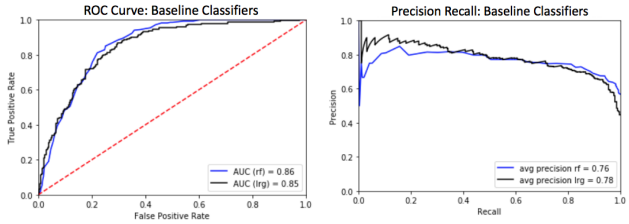



Figure 2: ROC curves with AUC in legend. Precision-Recall curves, with average precision scores in legend.

To evaluate the performance of the classifiers on our test data set, we first computed the average accuracy, average precision, and false negative rate, as shown in Table 1. We generated Receiver Operating Characteristic (ROC) curves and precision recall curves, as shown in Fig.6. The accuracy we received from simple baseline classification was surprisingly high—we found accuracy to be **76.3%** for random forest and **78.5%** for logistic regression. However, we found our false negative rate to be **9.2%** for random forest and **12.5%** for logistic regression. While we were initially impressed with the accuracy of these models, we realized that a false negative rate of around **10%** means that for one in ten patients, we would falsely diagnose that their melanoma was benign. This, as can be expected, can be disastrous for any sort of biomedical application and oncological in particular. Thus, our models urge to focus on decreasing this false negative rate, while also increasing accuracy. Table 1 summarizes all the information about classifiers performances.

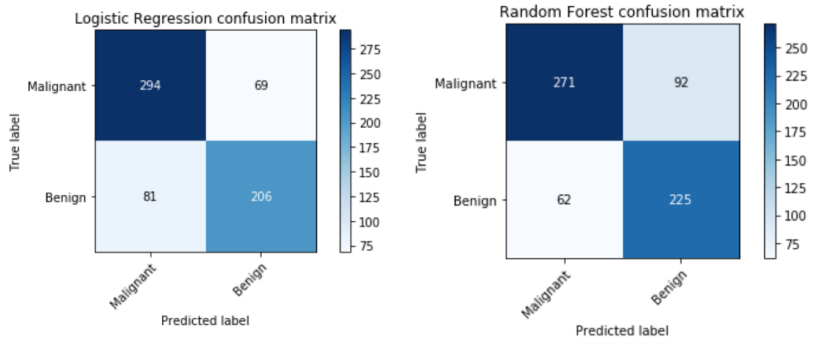



Figure 3: Confusion Matrices for Logistic Regression and Random Forest Classifiers

### 4.3 Convolutional Neural Network Classification Results

As per [7], we have implemented a ResNet, whose full specification is outlined in the Appendix at 2. We have chosen this model because it is not too deep and, given the fact that our dataset is relatively small when compared to the cardinality of DNN parameters, ResNet guarantees high performance rates as explained later in this section. The neural network described above was trained using the training set of 2600 images from the ISIC dataset. We began by training the CNN using 500 epochs as shown in Fig.4 and obtained an accuracy of **81%**, higher than either of the two baselines ( **76%**). When looking at the plot of model accuracy, we noticed the model overfitting to the data as shown by the validation loss increasing with an increase in the number of epochs. This lead us to look into early stop regularization, not allowing the CNN to continue if the validation loss continues to rise. As we see in Fig.5, this reduced the number of epochs to slightly over 100 and resulted in an improved accuracy of **87%**.

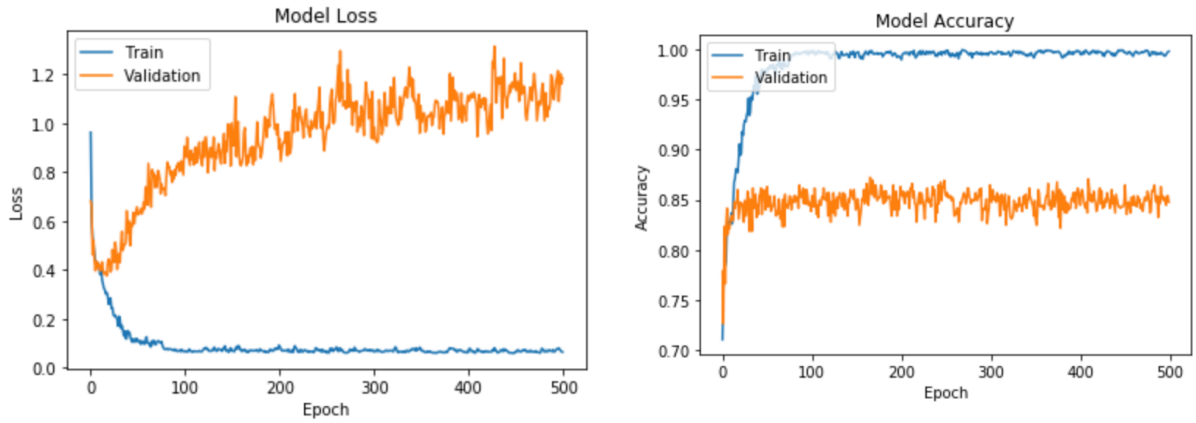



Figure 4: Model Loss and Model Accuracy for Convolutional Neural Network run with 500 epochs and no early stop regularization.

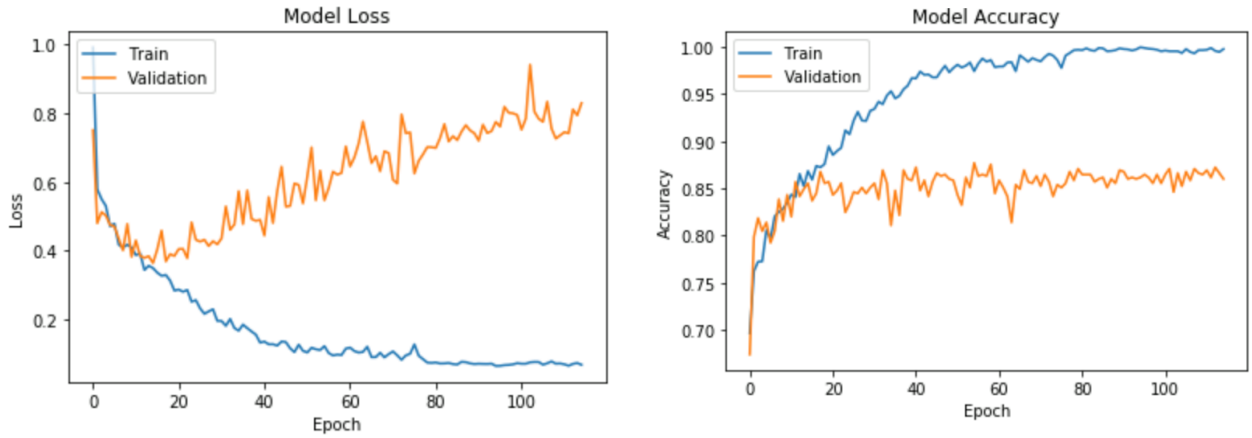



Figure 5: Model Loss and Model Accuracy for Convolutional Neural Network run with 100 epochs completed with early stop regularization

At the same time, as we can observe from the confusion matrix of below, the ResNet FNR (the fraction of melanoma erroneous diasgnoses) is **7.1%**, which is about a **5%** improvement from Logistic Regression and a **2%** one from Random Forest.

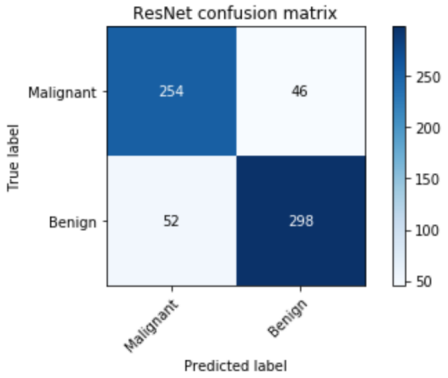



Figure 6: Confusion Matrix for ResNet Classifier.

As mentioned before, Table 1 summarizes all the information about classifiers performances.

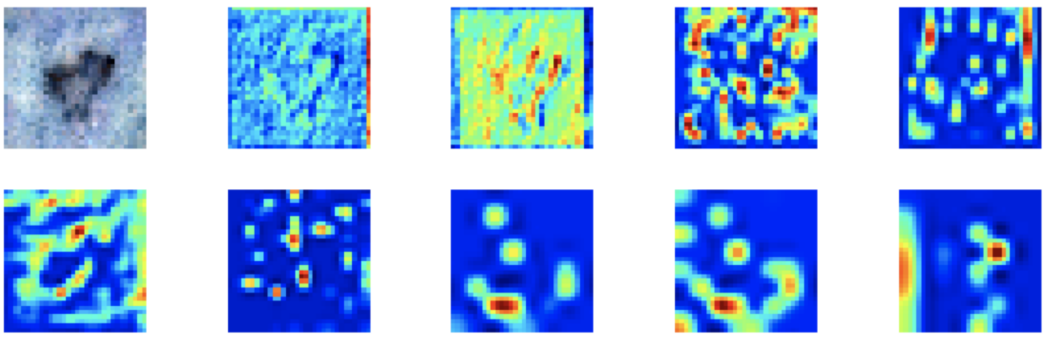
| Classifier | Accuracy | Precision | FPR | **FNR** |
|---|---|---|---|---|
| Logistic Regression | 0.769 | 0.785 | 0.106 | **0.125** |
| Random Forest | 0.763 | 0.764 | 0.144 | **0.092** |
| ResNet | 0.872 | 0.764 | 0.080 | **0.071** |

Table 1: Results from classifiers on test set with 80%-20% train-test split (2599 samples in training set and 650 in testing set).

This gives us confidence that the neural network is indeed a more powerful classifier than the baseline, however we are still left with the task of interpretation of the model as previously stated. For this we look at a correctly classified malignant image alongside the activation maps, LIME, and SHAP visualizations.

### 4.4 Comparison of CNN Visualizations

Fig. 7 shows each of the neural net visualizations on a correctly classified benign image from the testing set. Looking at the Class Activation Map, we can see the evolution of the convolutions and how they pick up on different important features of the image as described in Section 3.4. The first few layers look to be focusing on general features like shape and surroundings of the main center of mass. As the layers progress, areas of the lesion which are darker near the edge of the lesion are assigned more importance. This makes sense biologically as we will see in Section 4.5. Further confirmation of the oncological interpretation comes from the LIME and SHAP visualizations.

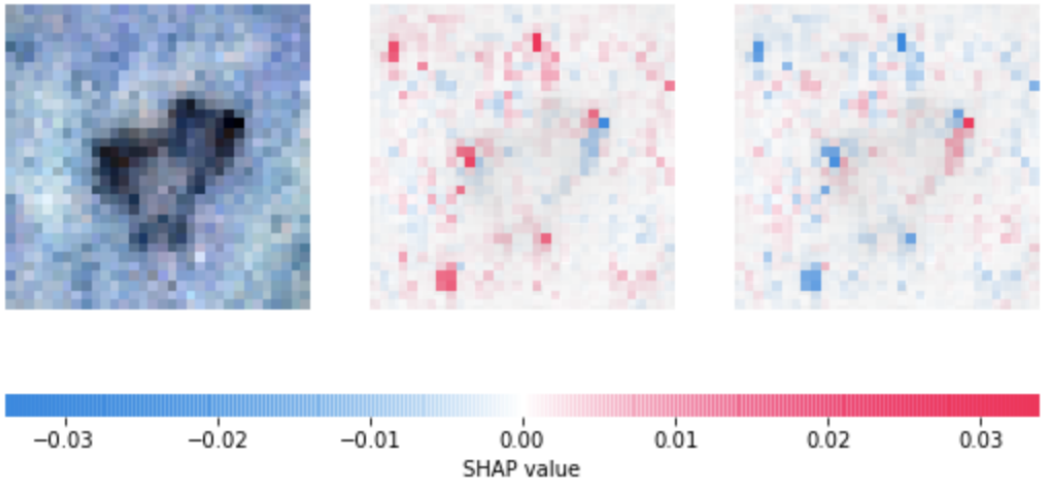


(a) Class Activation Map by Layer

Fig. 7b shows in its second and third image shows which part of the image explains respectively the benign and the malignant class. According to LIME, in fact, the lesion explains the malignant class which is what we would expect from a biomedical interpretation. Indeed, the image is correctly classified by the ResNet as malignant.


(b) LIME Visualization

On the other hand, according to SHAP, in Fig.7c) second image, we see that all the "hot" pixel values (in red) reside outside the lesion region of the image, further confirming that the benign class is explained by outside pixels: this implies that the boundaries are a really discriminatory feature (the SHAP last two images are one the color inverse of the other given that they explain complementary classes).




(c) SHAP Visualization

Figure 7: Correctly Classified Image

### 4.5 Biological Discussion

We were interested in comparing the areas of greatest weight in the CNNs classification with a visual diagnosis a melanoma lesion and see if there was significant overlap. The five key signs in visually diagnosing melanoma follow the ABCDE acronym: Asymmetry (half of the lesion does not match the other half), Border irregularity, Color (not uniform or featuring multiple colors), Diameter greater than 3mm, and Evolution of size, shape or color.

Figure 8: Full resolution image of correctly classified Melanoma image from Fig.7. Circles indicate areas in the original image with high SHAP values.

Interestingly, it appeared that there was not overwhelming overlap–the only major aspect that we noticed was that SHAP values appeared marginally higher around the lesion border, in the red circles displayed in Fig. 8. Its notable that the points with highest SHAP values also corresponded to areas of color contrast within the images, which could indicate the importance of color in the CNN's interpetation.

## 5 Discussion and Conclusion

From this project, we find reasonable accuracy from baseline classification methods of random forest and logistic regression, but find a much higher accuracy rate of **87%** with a Convolutional Neural Network. As we have mentioned, this accuracy comes at a bit of a cost with respect to interpretability given the non-linearity of the classifier. For this reason, we created visualizations of the model for our image data (Activation Maps, LIME and SHAP) in order to understand a few key features of the data. We see that the parameters of greatest variance are size and color of the lesion—reasonable findings given these two characteristics are two out of the five most important characteristics used to diagnose melanoma by human doctors.

This ultimately reinforces the desired relationship between machine learning and medical diagnosis. Deep learning techniques can be used for diagnosis with a very high accuracy, however the complementary function for human experts that comes along with the visualization may indeed prove more important—especially while deep learning techniques are still young as they are now. We can be optimistic that these methods will continue to evolve to improve medical diagnosis along with our understanding of medical processes at a deeper level.

# 6 Appendix

## 6.1 Dimensionality Reduction

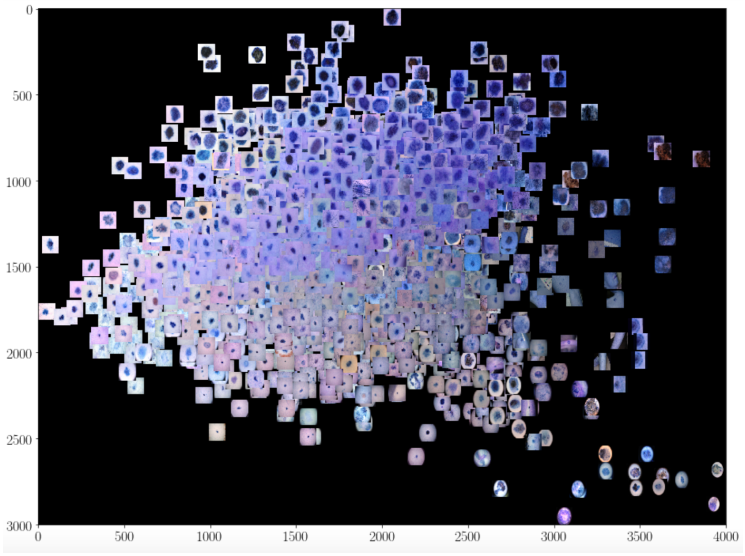



Figure 9: PCA of benign and malignant images.

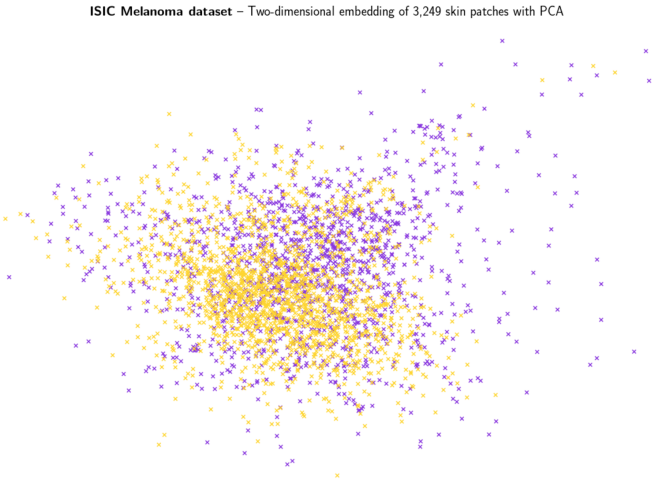



Figure 10: PCA on benign and malignant images. Yellow datapoints are benign, and black datapoints are malignant.

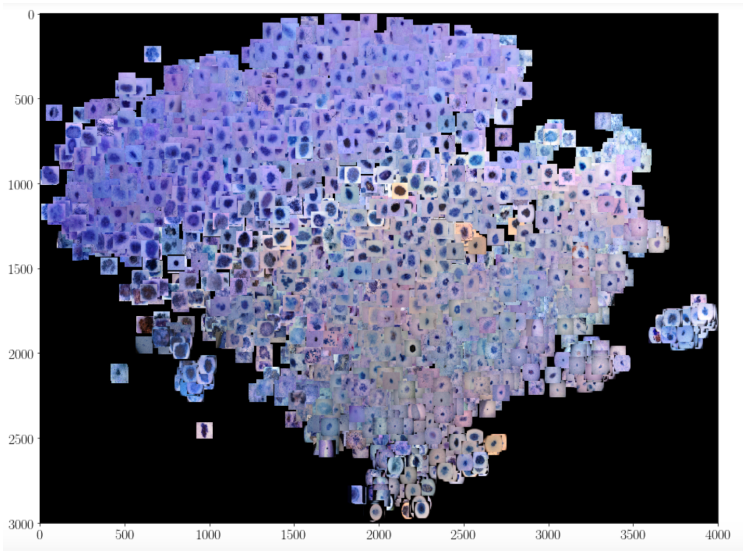



Figure 11: Visualization of class separation after performing t-SNE on PCA reduced dataset.

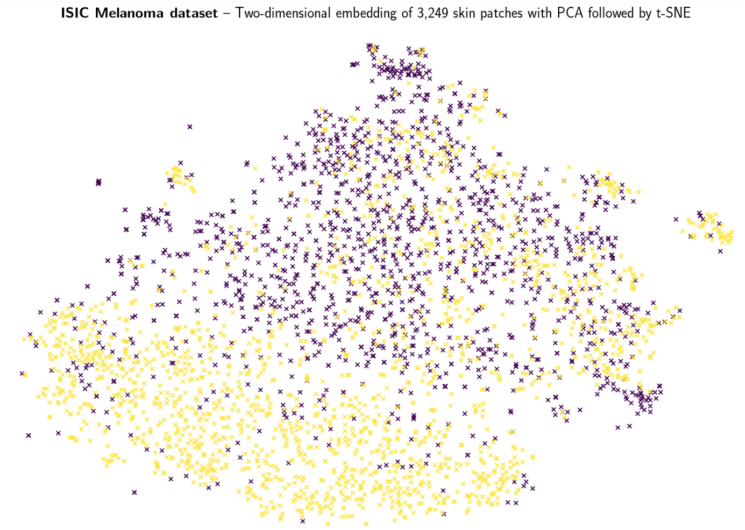



Figure 12: t-SNE on PCA reduced dataset of benign and malignant images. Yellow datapoints are benign, and black datapoints are malignant.

### 6.2 Deep Convolutional Neural Network Architecture

| Layer | Type | Output Shape | Parameters |
|---|---|---|---|
| conv2d_1 | Conv2D | (None, 32, 32, 32) | 896 |
| activation_1 | Activation | (None, 32, 32, 32) | 0 |
| batch_normalization$_1$ | Batch Norm | (None, 32, 32, 32) | 128 |
| conv2d_2 | Conv2D | (None, 32, 32, 32) | 9248 |
| activation_2 | Activation | (None, 32, 32, 32) | 0 |
| batch_normalization_2 | Batch Norm | (None, 32, 32, 32) | 128 |
| max_pooling2d_1 | Max Pooling | (None, 16, 16, 32) | 0 |
| dropout_1 | Dropout | (None, 16, 16, 32) | 0 |
| conv2d_3 | Conv2D | (None, 16, 16, 64) | 18496 |
| activation_3 | Activation | (None, 16, 16, 64) | 0 |
| batch_normalization_3 | Batch Norm | (None, 16, 16, 64) | 256 |
| conv2d_4 | Conv2D | (None, 16, 16, 64) | 36928 |
| activation_4 | Activation | (None, 16, 16, 64) | 0 |
| batch_normalization_4 | Batch Norm | (None, 16, 16, 64) | 256 |
| max_pooling2d_2 | Max Pooling | (None, 8, 8, 64) | 0 |
| dropout_2 | Dropout | (None, 8, 8, 64) | 0 |
| conv2d_5 | Conv2D | (None, 8, 8, 128) | 73856 |
| activation_5 | Activation | (None, 8, 8, 128) | 0 |
| batch_normalization_5 | Batch Norm | (None, 8, 8, 128) | 512 |
| conv2d_6 | Conv2D | (None, 8, 8, 128) | 147584 |
| activation_6 | Activation | (None, 8, 8, 128) | 0 |
| batch_normalization_6 | Batch Norm | (None, 8, 8, 128) | 512 |
| max_pooling2d_3 | Max Pooling | (None, 4, 4, 128) | 0 |
| dropout_3 | Dropout | (None, 4, 4, 128) | 0 |
| flatten_1 | Flatten | (None, 2048) | 0 |
| dense_1 | Dense | (None, 2) | 4098 |

Table 2: Full ResNet architecture specification

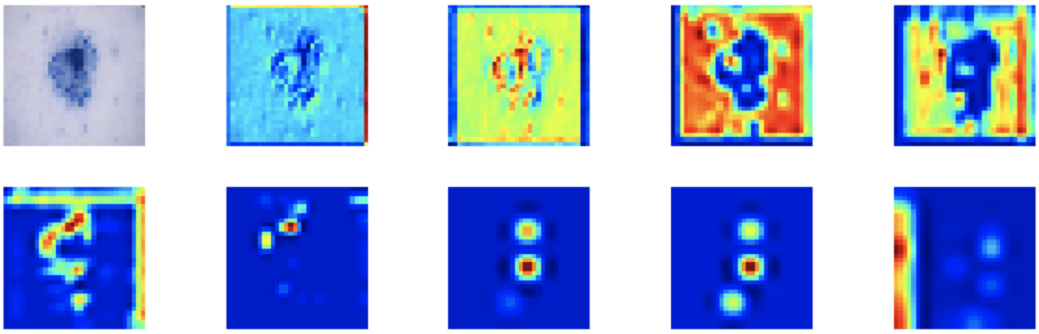
Thus, we have that:

- Total number of parameters: 292,898
- Trainable number of parameters: 292,002
- Non-trainable number of parameters: 896

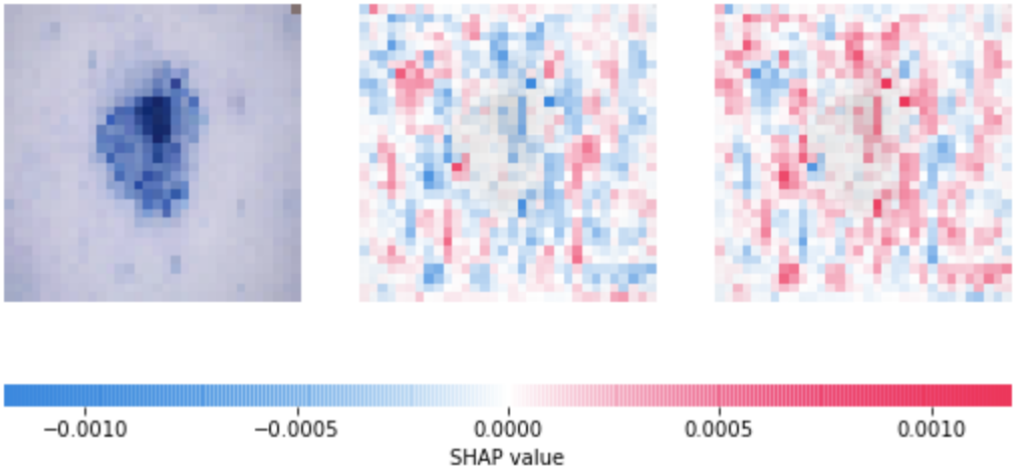
### 6.3 Interpretable Results in High-Confidence Misclassification Settings

We conduct the same type of analysis as in Section 4.4 but with respect to the following kind of image: the image has been misclassified with high degree of confidence. In practical terms, the model has predicted a benign tumour with high confidence, when that patient had, in fact, melanoma. We can observe how the activation maps across layers are not capable of putting under clear marking a melanoma lesion. We can also see how the ”hot” zones in the third activation map are at the exterior, meaning that not only it has detected the boundaries, but also is fooled into believing that the distinctive characteristic of that sample lie in the surroundings of the melanoma center of mass.


(a) LIME Visualization



(b) Class Activation Map by Layer




(c) SHAP Visualization

Figure 13: Misclassified Image

## References


[1] Different methods of image mapping, its advantages and disadvantages.

[2] International skin imaging collaboration: Melanoma project. 2017.

[3] Skin cancer. american academy of dermatology association.

[4] Key statistics for melanoma. *American Cancer Society*, 2018.

[5] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 2017.

[6] Dorina Kabakchieva. Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1):61–72, 2013.

[7] Abhijeet Kumar. Object-recognition-cifar-10. `https://github.com/abhijeet3922/Object-recognition-CIFAR-10`, 2018.

[8] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017.

[9] Philipp Fischer Olaf Ronneberger and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Springer Link*, 2015.

[10] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[11] Chiara Sabatti. Basic statistical references for the design and analysis of gene-chips experiments.

[12] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. *CoRR*, abs/1704.02685, 2017.

[13] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[14] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.