Melanoma diagnosis using deep learning techniques on dermatoscopic images

JACOB BROWN, UCSD, USA

This is a recreation of the paper of the same title [Jojoa Acosta 2021] by: Mario Fernando Jojoa Acosta1 , Liesle Yail Caballero Tovar1 , Maria Begonya Garcia-Zapirain1 and Winston Spencer Percybrooks. This paper implements a crop-then-classify model for detecting melanoma. I also compare various data augmentation techniques to increase balanced accuracy. I test the original techniques discussed in the paper and test two new techniques.

Additional Key Words and Phrases: Mask R_CNN, Deep learning, Transfer learning, Convolutional neural network, Object detection, Object classification

ACM Reference Format:
Jacob Brown. 2025. Melanoma diagnosis using deep learning techniques on dermatoscopic images. 1, 1 (June 2025), 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Motivation

The National Cancer Institute [National Institute of Health 2025] estimates that there will be 104,960 new melanoma cases and 8,340 deaths from melanoma in 2025. Melanoma is diagnosed primarily visually, as invasive procedures can damage the area and prevent tracking its evolution. Melanoma can spread to surrounding lymph nodes and metastasize to distant parts of the body[ACS 2023] so early detection is important. This makes the task perfect for image-based machine-learning models to classify skin lesions as benign or malignant. Dermatoscopic images, more specifically reflectance confocal microscopy, enable visualization from the epidermis all the way to the papillary dermis [Levine A 2018]. This enhanced imagery gives doctors a better understanding of the skin lesion and provides us with feature-rich images to train and detect. Computer evaluation of skin legions will vastly increase the number of skin lesions observed and can act as a screening tool for physicians. This will hopefully allow for melanoma to be detected at an earlier stage, reducing the damage it might do. Even with dermatoscopic images, there can be a considerable amount of noise preventing accurate diagnoses. Features such as hypo/hyper-pigmentation areas, inflammation, glints of light, colored patches, drops, hairs, ink, and oil around the affected area show the need for a special deep-learning architecture.

Author's Contact Information: Jacob Brown, UCSD, La Jolla, USA, jdb004@ucsd.org.

ACM XXXX-XXXX/2025/6-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Related Works

The International Skin Imaging Collaboration hosts a yearly competition for various tasks, and in 2017 and 2018 they hosted competitions centered around melanoma segmentation and detection. This competition resulted in many papers from participants that I will discuss. The "RECOD Titans at ISIC Challenge 2017" paper [Menegola et al. 2017], which ranked third, also performed image segmentation and classification as the paper I am replicating but they used three Inception-v4 models and one ResNet-101 model for classification combined with a support vector machine to get their final classification. This model scored fairly well in AUC and specificity measurements but was lacking in sensitivity and balanced accuracy. "Skin Lesion Segmentation in Dermoscopic Images With Ensemble Deep Learning Methods" [Goyal et al. 2020], also used the ISIC 2017 dataset with a similar method of a Mask-RCNN model but uses the predictions of both the Mask-RCNN and a DeepLabV3+ model as opposed to only a ResNet-152 model. They scored marginally better in sensitivity and specificity but did not report balanced accuracy and AUC measurements.

Going beyond the ISIC competition, "Skin cancer detection using ensemble of machine learning and deep learning techniques" [Tembhurne et al. 2023], uses a voting system between a VGG-16 CNN, a SVM, and a logistic regression. The images are preprocessed with principal component analysis for the SVM and logistic regression models. They achieve a better sensitivity score and accuracy than our model but it is unclear which ISIC dataset they used. "Detection of Malignant Melanoma Using Deep Learning" [Gulati and Bhogal 2019], compares AlexNet and VGG-16 on the PH^2 dataset which is considerably smaller than ISIC 2017. They found VGG-16 had better accuracy and sensitivity indicating larger models were still helpful in this problem. It is also important to note their test set was only 40 images. The paper, "Melanoma Detection Using Deep Learning-Based Classifications"[Alwakid et al. 2022], uses the HAM10000 dataset to train a variant of ResNet-50 with an extra fully connected layer and a custom CNN model. They found their model got better accuracy than the ResNet-50 but it is important to note they used provided masks to crop their image their model would need to be combined with a mask generation model to be deployed.

3 Project Aim

This project aims to solve the issue of noise in the classification of skin lesions and test various methods for improving the balanced accuracy of a binary classifier. Since this paper is a recreation of an existing paper I also aim to verify the results and replicability of their findings. I am also introducing two new solutions to improving balanced accuracy: weighting the loss function towards the malignant class and using a larger dataset. The larger dataset I tested with has a worse malignant to benign balance but I want to see if simply having five times more images total impacts the balanced accuracy.

4 Methodology

This paper proposes a two-stage classification model to first crop the image, remove the unnecessary features, and then classify the cropped image. The three datasets used are the ISIC 2017[Codella et al. 2017], ISIC 2018[Codella et al. 2019][Tschandl 2018], and the PH^2 [Teresa Mendonça 2013] datasets.

The Mask R-CNN model [He et al. 2018] first extracts features from the image with a backbone model, I used a ResNet 50 CNN for this. Then the model generates various bounding boxes in its region proposal network and then refines these proposals and sorts them by a probability score. The model also produces a mask as an output although the paper does not seem to use it so I do not include it in my training. The model also generates classifications but these are not used as we are using a much bigger ResNet-152 model to classify. The paper does not justify its design choice for using a Mask R-CNN over a Faster R-CNN since they do not use the outputted mask. I trained the model using losses for the region proposal network and the final bounding box output. Since we are only considering one lesion per image in this dataset I chose the final bounding box as the output with the best probability score. I generate the ground truth bounding boxes from the provided masks in the dataset. The paper did not specify if their Mask R-CNN model was trained from scratch or if they employed transfer learning like their ResNet model. To stay consistent with their ResNet model I employed a model trained on the COCO dataset provided by PyTorch and then applied transfer learning with the ISIC 2017 dataset. The original paper did not provide data augmentation results for the Mask R-CNN so I provided techniques one through four (to be discussed later) and excluded those that required new masks. For my loss function, I employed L1 loss (equation 1, since the loss function was not defined in this paper but this loss was consistent with other trainings of Mask R-CNN. I employed early stopping to save the model once it reached its lowest loss on the validation set to prevent overfitting. Sjöberg and Ljung [Sjöberg et al. 1995] argue that this is a form of implicit regularization to reduce the variance of the model.

$$Loss_{L1} = \sum_{i=1}^{n} |Target - Pred| \tag{1}$$

To implement the training, I needed to create a custom dataset class that could apply the needed transformations and provide the necessary labels. To train a Mask R-CNN model it needs a list of dictionaries with three elements: class, mask, and bounding box. These three elements also need to be tensors since there can be any number of labels per image (in our case there is only one). All these different data structures provide some additional difficulties in memory consumption and handling. Due to the size of my datasets, I am unable to train on UCSD's datahub servers, which have a maximum capacity of 10 GB so I train on my personal computer. I have 12 GB of VRAM on my GPU but when I tried to run a batch size of 4 for the Mask R-CNN model I quickly exceeded this memory capacity and incurred unreasonable performance penalties. To fix this I employed Torch's automatic mixed precision (AMP) mode. This automatically changes the bit precision of the floating points inside the model to the smallest possible size without reasonable loss of accuracy or risk of overflow. This mode allowed me to run a batch size of 8 using 11 GB and thus resulted in better speed and model performance. The use of automatic mixed precision also promotes the use of a gradient scaler to prevent vanishing or exploding gradients due to the limited precision. The nested data structure requested by the model also needed me to build a custom collate function for the data loader to properly pass the labels.

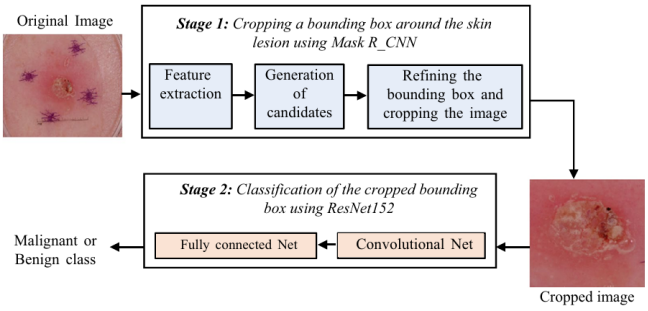

Fig. 1. Overall structure of the model. From the original paper [Jojoa Acosta 2021]

In the next stage, I use a ResNet 152-layer model [He et al. 2015] to classify the cropped image into either malignant (1.0) or benign (0.0). The benign classification represents anything that is not melanoma, thus it might miss other serious conditions. The malignant class identifies the skin lesion as being melanoma. I wanted to avoid using two models at the same time during training to keep my batch sizes high so I decided to precompute the bounding boxes for the ResNet training images. I made a Jupyter Notebook that runs the Mask R-CNN to output its predicated bounding boxes to a special csv and match each box with its respective image name.

The ResNet structure contains two major parts: the 152-layer convolutional neural network and a final fully-connected neural network. The ResNet model is one of the first convolutional neural networks to reach high accuracy with a very deep neural network. This performance is achieved with residual connections between the blocks that add the previous block's output into the next block's input. This helps to eliminate the vanishing gradient that previously dominated deep neural networks. Inadvertently this means that the model was similar to Euler's method for solving ordinary differential equations which also explains the model's efficacy. The ResNet model also includes batch normalization after each convolution so I wanted to keep the batch size as high as possible and thus I settled on a batch size of 64. The ISIC 2017 dataset contains only 1995 images for training which is quite small to train a large network and ResNet-152. To get around this the original paper employs transfer learning by loading a model already trained on the ImageNet-1K dataset. We load the model and weights provided by PyTorch and replace the final fully connected linear layer with a new layer with the same number of inputs but only two outputs to fit our binary classification. The loss function is not defined in the paper so I use a cross-entropy loss (equation 2 and test various weights to help bias towards the malignant class. I trained each model with a learning rate of 0.001 for 100 epochs.

Table 1. Data Augmentations and their respective ratios

Model	Modification to the malignant/benign training ratio	Malignant / Benign training ratio
Model 1	None	0.230
Model 2	Transformations applied to “malignant” class images	0.437
Model 3	Transformations applied to “malignant” class images, reduction “benign” class image	0.463
Model 4	Transformations applied to “malignant” class images, reduction “benign” class image	1.000
Model 5	Addition of PH2 images applied to “malignant” class images	0.254
Model 7	Only ISIC 2018 images are used	0.125

$$Loss_{CE} = Target * log(Pred) + (1 - Target) * log(1 - Pred) \quad (2)$$

For my data augmentation techniques, I employ 8 different techniques. Model 1 is my base model and model 2 is trained with 90% percent of the malignant images being duplicated and transformed to enlarge the dataset. The transformation I perform is a horizontal and vertical flip of the images. In model 3, I carry over model 2’s images but randomly remove benign images to increase the ratio of malignant to benign images. For model 4, I increased model 2’s augmentation to 100% and reduced the benign images until it was at a one-to-one ratio. For model 5, I include malignant images from the PH^2 dataset. Model 6 is obtained by retraining the best model for another 100 epochs at a training rate of 0.0001. Model 7 is obtained by training the Resnet with images from ISIC 2018 which has five times more images but a malignant to benign ratio half the size. Model 8 was obtained by training model 1 with different loss weightings, i.e. a malignant weighting of 0.6 to a benign weighting of 0.4 bears the model name "ResNet152_1_0.6". Models 5, 6, 7, and 8 are applied only to the ResNet training while 1, 2, 3, and 4 are applied to both Mask R-CNN and ResNet. The ratios are defined in Table 1

I computed various metrics to evaluate my model. Specificity, the performance of the benign class, and sensitivity, the performance of the malignant class, are commonly used. Accuracy is one of the most common metrics for evaluating classification models but when data is heavily biased towards one class, like in our datasets, it does not tell us how the model performs. For example, an accuracy of roughly 75% can be achieved by only guessing that the skin lesion is benign. In that example, a 75% accuracy is dangerous as untreated melanoma can have deadly consequences. The authors of the original paper use a balanced accuracy [Brodersen et al. 2010] to judge their model. The balanced accuracy that I use is defined as equation (3) since the original authors did not define a cost function for their balanced accuracy. I use balanced accuracy as the metric to determine the early stopping of the model’s training. This helps to regularize the model in a way that better represents both classes. I also compute the Reciever Operating Characteristic (ROC) of the model and I plot them all in a single graph. The closer to the top left of the graph a model is, the better it is, while also showing how the model balances true positives and false positives.

$$\frac{1}{2}(\frac{TP}{P} + \frac{TN}{N}) \quad (3)$$

Table 2. L1 Losses of all the Mask R-CNN Models

Model	Loss
Mask-RCNN_1	123.246941
Mask-RCNN_2	129.122910
Mask-RCNN_3	**112.198112**
Mask-RCNN_4	129.350555

My test setup is an Intel 10900k cPU with 32 GB of RAM paired with an Nvidia RTX 3080ti GPU with 12 GB of VRAM and the total size of the project was 43 GB, although that considers all the model storage and dataset duplication. It is important to state for others who wish to replicate these results that this computer was running at 100% usage during training so if you are planning to replicate this study make sure you have enough VRAM and storage. Better results may or may not be achievable if you can run higher batch sizes.

5 Results

In this section, I will show all of the results I achieved from each model and data transformation that I covered in the Methods section.

For Model 1, I trained it with just the ISIC 2017 dataset and its 1195 training images at a ratio of 0.230, 129 validation images, and 598 test images. I tested all models on this test dataset unless specified that it was trained on the ISIC 2018 dataset.

For the Mask R-CNN model, my only metric was L1 loss on the test dataset. The Mask R-CNN model was also only trained like the models one through four. In Table 2 we can see that loss does not meaningfully change by simply adding or removing images as adding malignant images and removing benign ones in model 3 reduces the loss while only adding malignant images or adding and subtracting more than model 3 both result in the same loss. These differences are more than likely caused by the inherent randomness when training a deep learning model.

To properly evaluate the ResNet Model I constructed Table 3 made up of all the metrics I calculated along with which dataset I tested them on. In Figure 2 I demonstrate the need to address the data imbalance as there are more false negatives than true positives which can have drastic consequences.

Model 2 shows considerable promise with 1% better accuracy, balanced accuracy, and precision while being 7% and 8% better on recall and F1 respectively. Observe in Figure 3 that our false negatives and true positives are closer together but this is still far

Table 3. Summary of All the Models and Their Metrics

Model	Test Dataset	Accuracy	Balanced Accuracy	Precision	Recall/Sensitivity	F1	Specificity
ResNet152_1_0.5	Dataset_2017	0.821667	0.711932	0.559524	0.401709	0.467662	0.923395
ResNet152_2_0.5	Dataset_2017	0.830000	0.728024	0.577320	0.478632	0.523364	0.915114
ResNet152_3_0.5	Dataset_2017	0.815000	0.705080	0.525862	0.521368	**0.523605**	0.886128
ResNet152_4_0.5	Dataset_2017	0.776667	0.669883	0.445860	**0.598291**	0.510949	0.819876
ResNet152_5_0.5	Dataset_2017	**0.835000**	**0.746628**	**0.632353**	0.367521	0.464865	**0.948240**
ResNet152_6_0.5	Dataset_2017	0.831667	0.732309	0.593023	0.435897	0.502463	0.927536
ResNet152_7_0.5	Dataset_2017	0.793333	0.620433	0.418605	0.153846	0.225000	**0.948240**
ResNet152_6_0.5	Dataset_2018	0.769180	0.605950	0.270619	**0.614035**	0.375671	0.788963
ResNet152_7_0.5	Dataset_2018	**0.902116**	**0.764685**	**0.603604**	0.391813	**0.475177**	**0.967189**

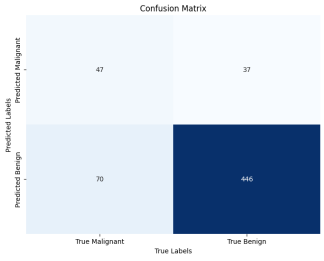

Fig. 2. Confusion Matrix of Model One

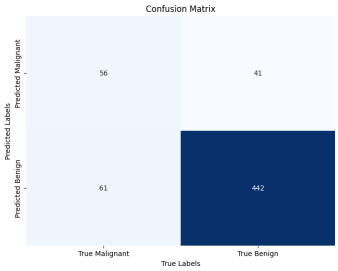

Fig. 3. Confusion Matrix: Model Two

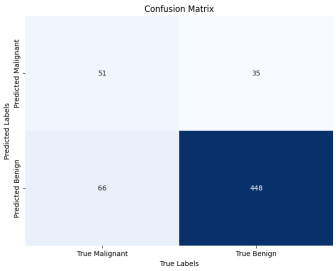

Fig. 4. Confusion Matrix: Model Six

from ideal. Still, this model has four more false positives than before but in our case that is preferable.

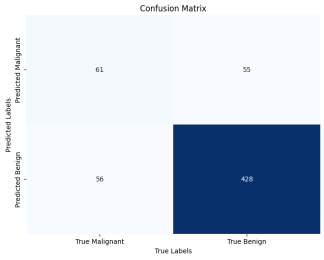

Fig. 5. Confusion Matrix of Model Three

Model 3 unfortunately performs worse on every metric compared to model 2 except where it does 5% better for recall and the same for F1. This tradeoff can be visualized in Figure 5 where we finally have more true positive guesses than false negatives, at the cost of more false positives.

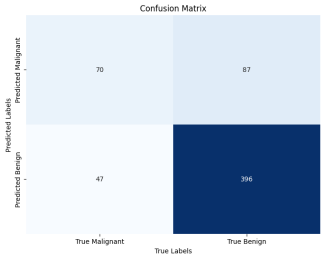

Fig. 6. Confusion Matrix of Model Four

Model 4 continues this trend by performing considerably worse on every metric except recall where it performs 7% better than model 3. In Figure 6 we can see that while our True positives have increased, our false positives grew so much they now outnumber the true positives. I have very little confidence in this model to make accurate predictions due to these ratios.

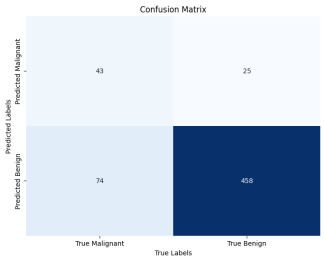

Fig. 7. Confusion Matrix of Model Five

Model 5 which trained on additional models from the PH^2 dataset performs the best of those testing on the ISIC 2017 dataset in all areas except for F1 and recall where it performs the second worst. My confusion matrix in Figure 7 demonstrates that this performance is achieved by focusing on the benign class as false negatives are significantly higher than even model 1 while false positives are the lowest among my models.

To align with the original paper my model 6 was a fine-tuned model and performed better on every metric compared to model 2 except recall and sensitivity which implies it started to prioritize the benign class. In Figure 4 we can see false negatives are increased as false positives decrease revealing our fine-tuning of model 2 provided the opposite of what I intended.

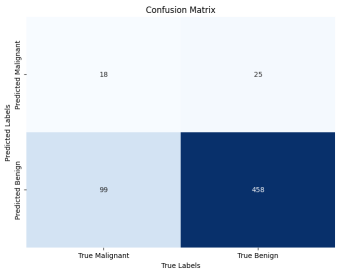

Fig. 8. Confusion Matrix: Model Seven (2017)

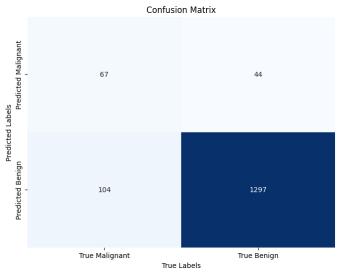

Fig. 9. Confusion Matrix: Model Seven (2018)

Training on the ISIC 2018 dataset, model 7, provided some interesting results. It has the highest scores in addition to mediocre recall and F! but only when tested on the 2018 test dataset. If model 7 is tested on the 2017 test dataset it has the lowest recall and F1 by 50% of the next lowest. Testing on the 2017 dataset further demonstrates the importance of data-balancing when training as, in Figure 8, it predicts lesions as malignant for only 43 images (only 18 of those were correct). Comparing this with Figure 9 we can see that its malignant predictions are more correct yet it still has a high false negative rate.

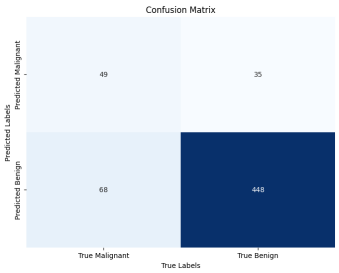

Fig. 10. Confusion Matrix: Loss weighting of 0.55

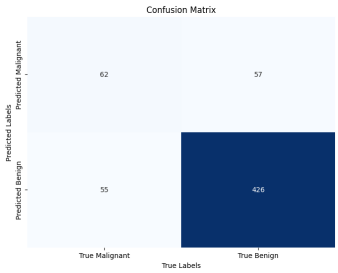

Fig. 11. Confusion Matrix: Loss weighting of 0.75

During training, I discovered I could apply weights to the loss to bias different classes so I decided to add that as a technique to help with the biased dataset. In Table 4 you can find the weights as the last value of the model name, for example, 0.6 indicates the loss of the malignant class is weighted 0.6 and the benign class is weighted 0.4. Accuracy, balanced accuracy, precision, and specificity reach a peak with a loss weighting of 0.55. F1 and recall peak at a loss weighting of 0.75, which is one minus the malignant to benign ratio. At 0.75 it has the highest F1 score out of all the models and the second highest recall only being passed by model 6 being tested on the ISIC 2018 dataset. From Figures 10 and 11 we can see that increasing the weighting of the malignant class effectively reduces false negatives while increasing false positives at a roughly equivalent weight.

Finally, I projected each of these models onto the Receiver Operating Characteristics (ROC) space, shown in Figure 12. The diagonal line in this figure represents a random guess curve since the bottom left of the space is guessing only negative and the top right is guessing only positive classifications. An ideal model reaches the top left of the space indicating a 100% correct classification of skin lesions. We can use this graph to visualize the performance trade-off of each model. An interesting result of this graph is that model 6, when tested on the ISIC 2018 dataset has the highest true positive rate while also having the highest false positive rate likely due to the composition of the 2018 dataset. Our next best contender was model 4 with similar True positives to model 5(2018) but with fewer false positives. Model 7, trained on the 2018 dataset, has the lowest false positives when tested on both datasets but has an incredibly low true positive rate. Our loss-weighted models seem to increase the true positive rate quite quickly until a weight of 0.8 where it matches the original model with more false positives. Model 6 (2017) offers slight improvements over the base model but model 2 seems to offer an even better balance.

6 Discussion

Further analyzing the results, I found that training on more images without regard to the distribution of the data worsens the performance of the model if the distribution is even more skewed. This is especially apparent when I trained a model on the ISIC 2018 dataset with five times as many images, it achieved a quite abysmal sensitivity rating. This is likely because the ratio of malignant to benign images is only half that of the 2017 dataset. Even adding only malignant images from a different dataset (model 5) had 10% worse recall than without. This aligns with the original paper's results and is probably due to different image processing techniques between datasets hindering proper feature extraction. Thus, when combining datasets, attention should be paid to pre-processing to ensure the datasets are compatible.

Another key finding I have found is that performing transformations on images (like in models 2 through 4), along with boosting the benign-to-malignant ratio, allows the model to better generalize to unseen data. When testing model 6 on the ISIC 2018 dataset, which contains considerably more images with different processing techniques, the model achieves the best sensitivity while having the worst specificity because the increase of malignant images from the transformations allows it to better recognize melanoma on new data. This aligns with the original dataset's findings but the difference in performance between models 1 and 2-4 is larger in the original paper.

My testing with loss weighting demonstrates that, when using skewed datasets, a small amount of bias towards the underrepresented class can slightly improve every metric I tested (Table 4). I also found that larger biases increase sensitivity (+32%) and F1 (+12%) significantly when compared to no biasing. This comes at the cost of overall/balanced accuracy (-1.1%,-1.3%), precision (-6.8%), and specificity(-4.5%). Going beyond a bias of 0.75 yields negative

Table 4. The Effects of Loss Weighting

Model	Test Dataset	Accuracy	Balanced Accuracy	Precision	Recall/Sensitivity	F1	Specificity
ResNet152_1_0.5	Dataset_2017	0.821667	0.711932	0.559524	0.401709	0.467662	0.923395
ResNet152_1_0.55	Dataset_2017	**0.828333**	**0.725775**	**0.583333**	0.418803	0.487562	0.927536
ResNet152_1_0.6	Dataset_2017	0.826667	0.722686	0.580247	0.401709	0.474747	**0.929607**
ResNet152_1_0.65	Dataset_2017	0.813333	0.701514	0.522124	0.504274	0.513043	0.888199
ResNet152_1_0.75	Dataset_2017	0.813333	0.703332	0.521008	**0.529915**	**0.525424**	0.881988
ResNet152_1_0.8	Dataset_2017	0.791667	0.662306	0.462264	0.418803	0.439462	0.881988

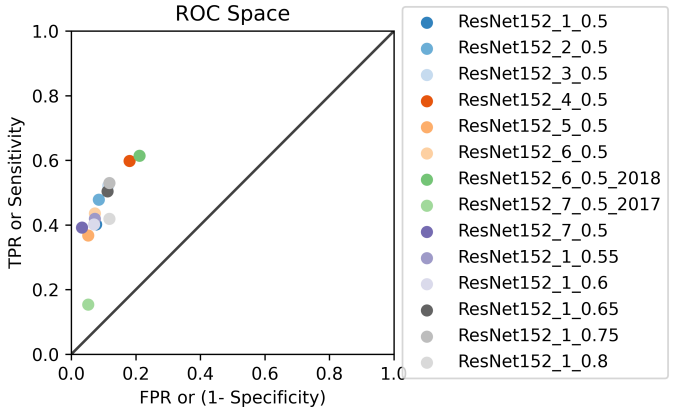

Fig. 12. Comparing models with the ROC space

changes across the board. Intuitively, this lines up with the distribution on the dataset (75% benign images) but this single example is not enough data to determine a relationship for the effectiveness of weighting the loss function. I believe it would be a promising avenue to combine their weighting with Model 2's augmentation of malignant images to create a better model.

In regards to the Mask R-CNN function, the original paper did not specify much information regarding its training and provided no metrics for it to be evaluated. The bounding box generation is a very important part of this model as if it fails the ResNet will be given an image with no skin lesion to classify. I observed this with my Mask R-CNN function when examining outputs from the test dataset and felt the original paper was lacking in its reproducibility due to their treatment of the Mask R-CNN model. In my paper I provide the L1 loss averaged across the test dataset as a way to compare this model against future reproductions (Table 2. I also trained this with data augmentation techniques 2-4 and used these models when training their respective ResNet models. I found no considerable relationship between the data augmentation techniques and their respective losses, thus I believe this is an open area for improvement of the proposed 2-stage model.

I found the core data augmentation techniques (models 2 to 4) each provided their own advantage over the base model and each other. Model 2 marginally improved upon the base model in every way except specificity (-0.9%). Similar to the original paper, this model provided the best overall score so I fine-tuned it (model 6). This had the opposite effect on the paper because this led the model to prioritize the benign class to achieve better loss. Model 3 improved on sensitivity but was weaker on every other metric (F! was equivalent to model 2). Since Model 4 is Model 3 with more aggressive benign image pruning, it follows the same trend with another decrease in every metric except for a significant increase in sensitivity. If you are primarily aiming for high sensitivity I would recommend models 3 and 4, otherwise, model 2 is the best choice.

Unfortunately, I was unable to achieve the performance of the original paper's model, eVida, by a significant margin shown in Table 5. I believe this to be potentially caused by a few reasons, mainly relating to a lack of information in the original paper. They did not specify the loss or optimization function they used to train either of their models, which can have a significant impact on a model's performance. They did not mention how well their Mask R-CNN model performed, so mine might have been skewing the results with incorrect image cropping. The weights for the balanced accuracy equation were also not given so I used an equal weighting to measure mine. For these reasons I find the paper's model results to be lacking in reproducibility although its findings on data augmentation techniques still proved mostly reproducible.

7 Conclusion

In this paper, I build a two-stage deep learning classifier to identify possible instances of melanoma. The first stage consisted of a Mask R-CNN model to generate bounding boxes so that I could crop the images to only highlight what is important and send that to the second stage. The second stage is a 152-layer ResNet model pre-trained

Table 5. Comparison of this Paper and The Original Paper

Model	Accuracy	Balanced Accuracy	Recall/Sensitivity	Specificity
ResNet152_2_0.5	0.830	0.728	0.478	0.915
eVida	**0.904**	**0.872**	**0.820**	**0.925**

on ImageNet-1k dataset and I performed transfer learning on it with the ISIC 2017 dataset. I then tested various data augmentation techniques that included transforming malignant images, removing benign images, adding malignant images from a different dataset (PH^2), and training on a much larger dataset (ISIC 2018). I also evaluated the performance impacts of weighting the loss towards the malignant classification and found small amounts to be generally helpful and larger amounts to help with sensitivity. The main takeaway from this paper is that data augmentation can provide an effective avenue for increasing performance on unbalanced datasets.

Here is the link to my GitHub Project Repository

References

American Cancer Society ACS. 2023. *Melanoma Skin Cancer Stages.* https://www.cancer.org/cancer/types/melanoma-skin-cancer/detection-diagnosis-staging/melanoma-skin-cancer-stages.html

Ghadah Alwakid, Walaa Gouda, Mamoona Humayun, and Najm Us Sama. 2022. Melanoma Detection Using Deep Learning-Based Classifications. *Healthcare* 10, 12 (2022). doi:10.3390/healthcare10122481

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M. Buhmann. 2010. The Balanced Accuracy and Its Posterior Distribution. In *2010 20th International Conference on Pattern Recognition*. 3121–3124. doi:10.1109/ICPR.2010.764

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern. 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1902.03368 [cs.CV] https://arxiv.org/abs/1902.03368

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. 2017. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). arXiv:1710.05006 [cs.CV] https://arxiv.org/abs/1710.05006

Manu Goyal, Amanda Oakley, Priyanka Bansal, Darren Dancey, and Moi Hoon Yap. 2020. Skin Lesion Segmentation in Dermoscopic Images With Ensemble Deep Learning Methods. *IEEE Access* 8 (2020), 4171–4181. doi:10.1109/ACCESS.2019.2960504

Savy Gulati and Rosepreet Kaur Bhogal. 2019. Detection of Malignant Melanoma Using Deep Learning. In *Advances in Computing and Data Sciences*, Mayank Singh, P.K. Gupta, Vipin Tyagi, Jan Flusser, Tuncer Ören, and Rekha Kashyap (Eds.). Springer Singapore, Singapore, 312–325.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2018. Mask R-CNN. arXiv:1703.06870 [cs.CV] https://arxiv.org/abs/1703.06870

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV] https://arxiv.org/abs/1512.03385

Caballero Tovar Liesle Yail Garcia-Zapirain Maria Begonya Percybrooks Winston Spencer Jojoa Acosta, Mario Fernando. 2021. Melanoma diagnosis using deep learning techniques on dermatoscopic images. *BMC Med Imaging* 21, 6 (Jan. 2021). doi:10.1186/s12880-020-00534-8

Markowitz O. Levine A. 2018. Introduction to reflectance confocal microscopy and its use in clinical practice. *JAAD case reports* 4, 10 (Nov 2018), 1014–1023. doi:10.1016/j.jdcr.2018.09.019

Afonso Menegola, Julia Tavares, Michel Fornaciali, Lin Tzy Li, Sandra Avila, and Eduardo Valle. 2017. RECOD Titans at ISIC Challenge 2017. arXiv:1703.04819 [cs.CV] https://arxiv.org/abs/1703.04819

National Cancer Institute National Institute of Health. 2025. *Cancer Stat Facts: Melanoma of the Skin.* https://seer.cancer.gov/statfacts/html/melan.html

Jonas Sjöberg, Qinghua Zhang, Lennart Ljung, Albert Benveniste, Bernard Delyon, Pierre-Yves Glorennec, Håkan Hjalmarsson, and Anatoli Juditsky. 1995. Nonlinear black-box modeling in system identification: a unified overview. *Automatica* 31, 12 (1995), 1691–1724. doi:10.1016/0005-1098(95)00120-8 Trends in System Identification.

Jitendra V. Tembhurne, Nachiketa Hebbar, Hemprasad Y. Patil, and Tausif Diwan. 2023. Skin cancer detection using ensemble of machine learning and Deep Learning Techniques. *Multimedia Tools and Applications* 82, 18 (Feb 2023), 27501–27524. doi:10.1007/s11042-023-14697-3

Jorge Marques Andre R. S. Marcal Jorge Rozeira Teresa Mendonça, Pedro M. Ferreira. 2013. PH² - A dermoscopic image database for research and benchmarking. https://www.fc.up.pt/addi/ph2%20database.html [Accessed 07-06-2025].

Rosendahl C. Kittler H. Tschandl, P. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions - Scientific Data — nature.com. https://www.nature.com/articles/sdata2018161#citeas [Accessed 07-06-2025].