Image Based Melanoma Detection

Vyom Pathak*, Sanjana Rao†, Ira Harmon‡ and Daisy Wang§
Department of Computer & Information Science & Engineering, University of Florida
Gainesville, Florida, United States
{*v.pathak, †sanjanar.guttalu, ‡iharmon1, §daisyw}@ufl.edu

***Abstract*—We present an ensemble of image-only convolutions neural network (CNN) models with different backbones and input sizes along with a self-supervised model to classify skin lesions. We have devised the first ensemble based on the winning solution to Kaggle's SIIM-ISIC Melanoma Classification challenge (We will be referring to this as the Ha-CNN model going further) and Bootstrap your own latent (BYOL) model which is based on self-supervised learning. The models have experimented with the SIIM-ISIC Melanoma dataset (2018-2020). Using specificity and sensitivity as the performance metrics, nine top-performing models were selected out of the eighteen models proposed in the Ha-CNN paper. We experimented BYOL model with two different backbones - ResNet and EfficientNet. The Ha-CNN model achieves a specificity and sensitivity of 94.3% and 92.1% with a negative predictive value of 99.2. As with the BYOL model, our results show an increase of 1.00% for the ResNet-101 model supervision (94.73% and 93.40%) and an increase of 1.00% for the Efficient-B5 model (97.24% and 96.34%) with and without BYOL-self-supervision.**

***Index Terms*—melanoma, image classification, CNN, self-supervision, BYOL**

I. Introduction

Skin cancer has been one of the most deadly and the most common form of cancer occurring in the world [8]. The WHO estimated in 2003 that between two and three million skin malignancies develop globally each year, with basal cell carcinoma accounting for 80%, cutaneous squamous cell carcinoma accounting for 16%, and melanoma accounting for 4% (approximately 130,000 cancers) [9]. Despite its lower occurrence, melanoma skin lesions metastasize to other parts of the body very quickly meaning that it is responsible for up to 75% of skin cancer deaths [9]. However, early detection of melanoma has been linked with a five-year survival rate improvement of about 91-95%. As a result, fast and precise diagnostic tests can aid general practitioners and dermatologists in avoiding delays in disease treatment.

Several methods have been developed to detect melanoma. The Gold standards include clinical examination with biopsy or by the naked eye. Dermatologists and medical practitioners formally trained in different dermoscopic algorithms showed an average sensitivity for detecting melanoma of mostly $< 80\%$ [7]. Moreover, dermoscopy significantly improves the accuracy of the naked eye examination [19]. The downside of these methods is the need for very skilled dermatologists to perform the biopsy or the naked eye test. [14].

Epidermal genetic information retrieval (EGIR) methods also show good classification scores where adhesive lifts the epidermal cells which are used to perform genetic testing to detect melanoma [26]. These methods improve upon the biopsy method achieving a higher sensitivity (88%) by using the genetic information from the cells. This method still involves a high latency in results as well as requires expensive instruments for gene sequencing. This process can be still automated effectively detecting Melanoma from an image on a cell phone.

Recent advancements in machine learning more specifically deep learning have paved a path in developing image-based melanoma detection models. Here, high-resolution images of skin are classified as malignant or benign using Convolutional Neural Networks [14], [22]. One such CNN model based on Google's Inception v4 [20] developed by Haenssle et al. in 2018 outperformed the dermatologists in melanoma detection [14]. This system outperformed 58 dermatologists amongst which were 30 experts; by giving an AUC score of 78.16% on a subset of ISIC 2016 dataset [11]. With the launch of the SIIM-ISIC 2019-2020 challenge on Kaggle [4], [5] we see a CNN with meta-data based ensemble technique (18 models) which showed an AUC score of 94.9% [12]. We experiment with this technique to include models which do not use meta-data and select an ensemble of 8 models. For measuring the prediction score, we find the best threshold by maximizing the G-Means score [1] which proves to be a good measure in the melanoma classification task. For this task, because of the low amount of negative examples, it becomes difficult to develop a generalized algorithm that works under different scenarios. We achieved an NPV score of around 99.2% over a subset of the SIIM-ISIC 2019-2020 dataset.

Furthermore, there has been the development of self-supervised image representation learning techniques [3], [15], [23]–[25] that can take on this problem. Particularly Grill et. al. introduced Bootstrap Your Own Latent (BYOL) [10] which relies on two neural networks referred to as online and target networks which learn feature representations from each other. This technique does not rely on negative pairs for self-supervision making it a good candidate for the melanoma task. For our task, we first take a CNN architecture and perform self-supervision training using BYOL. Then, we further finetune this CNN architecture as a supervision task of melanoma classification. We performed 2 experiments on BYOL where we compared the BYOL performance over two different backbone architectures namely ResNet-101 [16] and EfficientNet-B5 [21]. We also compared the performance of self-supervision by comparing the performance of the model which is finetuned without self-supervision and with

supervision. We used the SIIM-ISIC dataset from the years 2019 and 2020. For metric calculation, we calculated the best threshold by maximizing the G-Means score similar to the CNN-Ha experiment. We see that efficientnetB5 shows a performance increase over the resnet model of about 3% in terms of the NPV value. Furthermore, we also show that BYOL supervised model performs better than the normal finetuned model showing an increase of around 1% for the resnet models and around 1% for the efficientnet-B5 model for a small subset of the SIIM-ISIC 2019-2020 dataset.

In section II, we provide a brief review of the related works that emphasize our proposed methods. Dataset details, model architecture, and experiment setup have been described in section III. We then compare our experimental results with established standard methods and further discuss any nuances observed from the results along with future work in section IV. Finally, we conclude the paper with future work in section V

II. Literature Review

Melanoma is the most deadly form of skin cancer, but it can be cured with minor surgery if caught early enough. Quick and accurate diagnosis could immensely benefit doctors and patients. Recent advancements in deep learning-based computer vision have pushed model performance to be close to (or exceed in some cases) the human expert level in many medical fields. The Ha-CNN model is based on the ensemble proposed by Ha et. al. [13]. The model achieves an AUC of 0.9490 by using a diverse group of state-of-the-art models trained with various input sizes and a good validation strategy. Another unique strategy used in the paper is the usage of metadata along with the images. However, we have not considered metadata to train our ensemble as we aim at using the ensemble as a service for a cell phone application.

There have been several attempts at developing self-supervised models for learning image representations, which can be divided as generative and discriminative techniques [3], [6]. Generative techniques typically operate directly in pixel scale. This is computationally expensive, and unnecessary to learn representation learning [10].

Amongst the Discriminative methods, contrastive learning achieves SOTA performance in self-supervised representation learning [15], [23]–[25]. However, contrastive learning requires comparison with a plethora of examples predominantly negative pairs. However, it becomes difficult to get a large number of negative pairs for the melanoma task. Thus, we look at Bootstrap Your Own Latent (BYOL) which learns image representations using self-supervision by using bootstrapping over its target network [10]. To further prevent collapse while representation learning, BYOL uses predictors on top of the online network which are compared with the target representations by using mean squared errors [10]. Our work is similar to Chaves [2], where they study the application of self-supervision for skin lesion detection for different training data scenarios as well as different self-supervision schemes. Our work is different in the sense that we compare the use of the EfficientNet-B5 model as opposed to the ResNet-101 model while emphasizing the improvement of the model performance for both of them as a result of self-supervision.

III. Method and Dataset Details

A. Dataset

The SIIM-ISIC 2020 dataset comprises 1.76% percent of the images which are malignant and hence lead to an unstable AUC score [4]. To overcome this, we have incorporated the years - 2019 [5] and 2020 datasets which yields an overall 8.7% of the images to be malignant. Our training data comprises a total of 58,437 images with input resolutions - 512, 768, and 1024. The test set consists of 10,982 images from the 2020 dataset. The dataset and code with the implementation step are available on github [1].

B. CNN-Ensemble

The Ha-CNN model comprises eighteen models. The first sixteen models use EfficientNet (B3, B4, B5, B6, B7) as the backbone and the last two models use Resnet (SE-ResNext-101, ResNeSt-101). As seen in the Fig. 1, before we train the models, we resize and perform data augmentations using the augmentation library Albumentations: Transpose, Flip, Rotate, RandomBrightness, RandomContrast, MotionBlur, MedianBlur, GaussianBlur, GaussNoise, OpticalDistortion, GridDistortion, ElasticTransform, CLAHE, HueSaturationValue, ShiftScaleRotate, Cutout to prevent over-fitting. Thereafter, We perform five-fold cross validation and use bagging to improve performance.

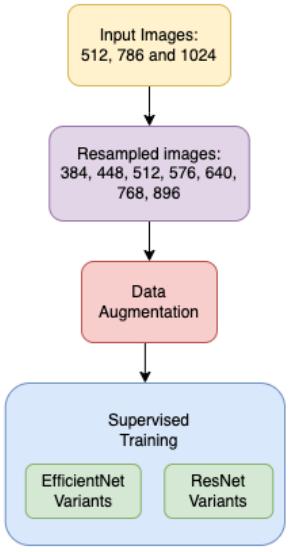

Fig. 1. Model Architecture for CNN-Ha experiment

The training setup includes a batch size of 64 for all models. The first model runs for 18 epochs while the rest are run for 15 epochs. Adam optimizer has been used with cosine annealing

[1]https://github.com/01-vyom/melanoma-classification

with one warm-up restart. Initial learning rates are tuned for each model and at the warm-up epoch (epoch 4), the learning rate is set to one-tenth the initial learning rate of the cosine cycle. We use cross-entropy as the loss function.

After sampling images equally from each of the five folds for each of the models, we computed the performance metrics for each of the models. The same was compared to select nine top-performing models from the set of eighteen models. We refrained from using the models which made use of the meta-data. Fig. 2 shows the ROC curve and top threshold values obtained for model-3.

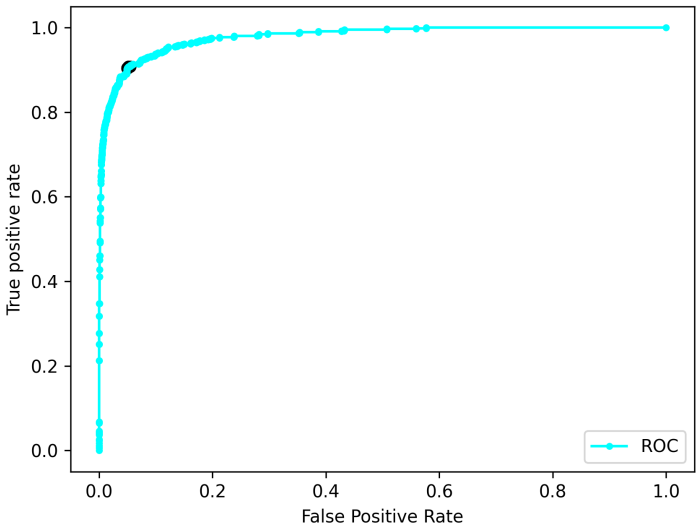

Fig. 2. ROC Curve for Model-3 with best threshold value of 0.076472

C. BYOL-CNN

The BYOL experiment comprises 2 different types of models. The whole architectural process is shown in Fig. 3. The first model uses ResNet-101 as a backbone architecture while the second model uses the EfficientNet-B5. Each of the backbone architecture is trained using self-supervision by passing it to the BYOL model. The BYOL model uses online and target networks to learn the image representation [10]. At a high level, the online network tries to predict an image that is generated by the target model using some augmentation functions for each of the network [10]. The online network uses the image projection to perform prediction, and the target network only defines the projection [10]. Then, the normalized predictions and target projections are compared using mean squared error [10]. This step helps us to learn the image representation. Then, we can use this self-supervised backbone to perform classification by finetuning it over labeled images. For each experiment, we compare the performance of self-supervision by training the CNN model with and without BYOL-self-supervision.

We used the SIIM-ISIC 2019 and 2020 datasets by splitting them into 80 and 20 ratios for training and validation for both self-supervision and supervision steps. We used 512-by-512 image resolution for supervision and a 448 self-supervision step. The validation dataset is also used for testing the supervision measure of the model. For training, we use a batch size of 32 for the ResNet-101 model and 25 for EfficientNet-B5. We use the Adam optimizer with a learning rate of $1e^{-4}$. We train the ResNet-101 model for 10 epochs for Supervision and Self-Supervision and EfficientNet-B5 for 20 epochs for Self-Supervision and 10 epochs for Supervision. The self-supervision step also considers using gradient accumulation.

For the testing phase, we test both the CNN-backbone architecture with and without self-supervision to compare the performance of the BYOL model. We use the 20% dataset to perform the testing. We calculate different threshold values from the AUC-ROC curve. Then, we find the best threshold by maximizing over the G-Means value. We then use this threshold to find other metrics.

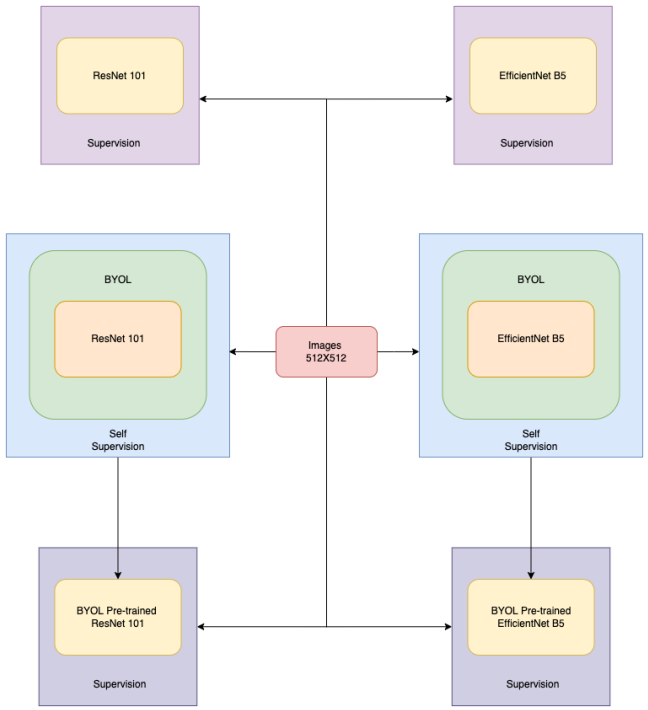

Fig. 3. Model Architecture for BYOL Experiment.

IV. Results

A. CNN-Ensemble

For each model, we computed the specificity and sensitivity by plotting a ROC curve to find the threshold values. We decided on the best threshold value by maximizing over the G-means value. These metrics were used to select the top-performing models.

The performance of the ensemble is a simple average of all models' probability. The model achieved an overall specificity of 94.3% and sensitivity of 92.1% with G-means of 93.3% and NPV value of 99.2%. The model achieves an AUC score of 98.2% as shown in Table I.

TABLE I
RESULTS

Model Name	Sensitivity(%)	Sensitivity(%)	G-mean	NPV	AUC
Dermatologists (Henselle [14])	88.9 ± 9.6	75.7 ± 11.7	82.0	-	82
EGIR [26]	88	100	93.8	> 99	95.5
CNN-Thissen [22]	78	80	78.9	-	-
CNN-Haenssle [14]	-	82.5 (@88.9 sen.)	85.6	-	95.3
CNN-Ha (original [12])	-	-	-	-	94.9
CNN-Ha (experiment)	92.1	94.3	93.3	99.2	98.2
BYOL - ResNet101	42.81	97.94	64.75	94.73	—
Resnet101	27.05	98.44	51.6	93.40	—
BYOL - EfficientnetB5	59.16	98.44	76.3	97.24	—
EfficientnetB5	60.35	98.40	77.1	96.34	—

B. BYOL-CNN

Table I shows the melanoma classification metrics for the ResNet-101 as well as the EfficientNet-B5 model. The metrics are calculated by finding the best threshold by maximizing the G-Means value. The threshold value for the ResNet-101 experiment is 0.0946 and 0.0077 for supervised classification with and without BYOL supervision. The threshold value for the EfficientNet-B5 experiment is 0.0040 and 0.0004 for supervised classification with and without BYOL supervision.

We observe after the CNN-Ensemble experiments that the ensemble performs very well and achieves a negative predictive value of 99.2. G-means prove to be a good measure to select the best threshold. Our results show an increase of 1.33% for the ResNet-101 model supervision (94.73% and 93.40%), and an increase of 0.90% for the Efficient-B5 model (97.24% and 96.34%) with and without BYOL-self-supervision. We can further improve the BYOL performance by using a 5 fold cross-validation strategy.

In future work, we want to build and train the Laina model [17] on the NYU Depth dataset [18] before integrating it with the BYOL model. We would next like to fine-tune it using integumed 3D data and concatenate it with chosen models from the CNN-Ha ensemble. This would then be processed by a linear layer to create predictions. The ensemble can then be added to any cell phone for fast melanoma classification.

V. CONCLUSION

In this paper, we build models to identify melanoma using images of skin lesions. We use several models with EfficientNet and ResNet backbones and use data augmentations. We make use of the available SIIM-ISIC datasets over the previous years to strengthen our processing and improve the model's learning. We evaluate and select models based on their specificity and sensitivity maximized over G-means. We obtain an overall AUC score of 98.2% and NPV value of 99.20% for the CNN-Ensemble experiment. For the BYOL experiment, we observe an increase of 1.33% for the ResNet-101 model supervision (94.73% and 93.40%) and an increase of 0.90% for the Efficient-B5 model with and without BYOL (97.24% and 96.34%).

In the future, we plan to develop and train the Laina model on NYU Depth dataset and then, integrate it with the BYOL model. We then would like to fine-tune the same using integumed 3D data and then concatenate it with selected models from the Ha-CNN ensemble. This would then be passed through a linear layer to make predictions. The ensemble model may then be utilized as a service for any smartphone to classify melanoma quickly.

REFERENCES

[1] Josephine Akosa. Predictive accuracy: A misleading performance measure for highly imbalanced data. In *Proceedings of the SAS Global Forum*, volume 12, 2017.

[2] Levy Chaves, Alceu Bissoto, Eduardo Valle, and Sandra Avila. An evaluation of self-supervised pre-training for skin-lesion analysis. *arXiv preprint arXiv:2106.09229*, 2021.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[4] International Skin Imaging Collaboration et al. Siim-isic 2020 challenge dataset. *International Skin Imaging Collaboration*, 2020.

[5] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

[6] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[7] Con Dolianitis, John Kelly, Rory Wolfe, and Pamela Simpson. Comparative performance of 4 dermoscopic algorithms by nonexperts for the diagnosis of melanocytic lesions. *Archives of dermatology*, 141(8):1008–1014, 2005.

[8] J Ferlay, M Colombet, I Soerjomataram, T Dyba, G Randi, M Bettio, A Gavin, O Visser, and F Bray. Cancer incidence and mortality patterns in europe: Estimates for 40 countries and 25 major cancers in 2018. *European journal of cancer*, 103:356–387, 2018.

[9] Karoline Freeman, Jacqueline Dinnes, Naomi Chuchu, Yemisi Takwoingi, Sue E Bayliss, Rubeta N Matin, Abhilash Jain, Fiona M Walter, Hywel C Williams, and Jonathan J Deeks. Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. *bmj*, 368, 2020.

[10] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

[11] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016.

[12] Qishen Ha, Bo Liu, and Fuxu Liu. Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge. *arXiv preprint arXiv:2010.05351*, 2020.

[13] Qishen Ha, Bo Liu, and Fuxu Liu. Identifying melanoma images using efficientnet ensemble: Winning solution to the siim-isic melanoma classification challenge. 2020.

[14] Holger A Haenssle, Christine Fink, Roland Schneiderbauer, Ferdinand Toberer, Timo Buhl, Andreas Blum, A Kalloo, A Ben Hadj Hassen, Luc Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology*, 29(8):1836–1842, 2018.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.

[18] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[19] G Salerni, T Terán, S Puig, J Malvehy, I Zalaudek, G Argenziano, and H Kittler. Meta-analysis of digital dermoscopy follow-up of melanocytic skin lesions: a study on behalf of the international dermoscopy society. *Journal of the European Academy of Dermatology and Venereology*, 27(7):805–814, 2013.

[20] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[21] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[22] Monique Thissen, Andreea Udrea, Michelle Hacking, Tanja von Braunmuehl, and Thomas Ruzicka. mhealth app for risk assessment of pigmented and nonpigmented skin lesions—a study on sensitivity and specificity in detecting malignancy. *Telemedicine and e-Health*, 23(12):948–954, 2017.

[23] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[24] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020.

[25] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018.

[26] W Wachsman, V Morhenn, T Palmer, L Walls, T Hata, J Zalla, R Scheinberg, H Sofen, S Mraz, K Gross, et al. Noninvasive genomic detection of melanoma. *British Journal of Dermatology*, 164(4):797–806, 2011.