Melanoma Diagnosis using Convolutional Neural Networks

Ansh Chaurasia *

February 2021

1 Introduction

A recent report by the World Health Organization (WHO) has included cancer in the top 10 causes of death [3]. More alarmingly, data from the same report also indicates the rate of patients diagnosed with cancer may double [9]. Cancer can be lethal to patients, but its effects can be mitigated if detected and treated early [7]. Therefore, it is worthwhile to invest time and research to improve our ability to diagnose cancers as early as possible.

Melanoma is one of the most lethal forms of skin cancer. It occurs in cells known as melanocytes, skin cells in the upper layer of skin. Melanocytes produce a pigment known as melanin to give the skin it's color. However, when skin is exposed to UV radiation, melanocytes produce more melanin than necessary, causing skin damage. Melanoma occurs when UV radiation causes mutations in these melanocytes, which leads to unrestrained cellular growth. Figure 1. gives a visual difference between benign and malignant melanoma images. By the end of 2020, about 196,060 people in the USA will be diagnosed with melanoma, and of these more than 100,000 people are expected to be diagnosed with invasive (penetrating the epidermis into the skin's second layer, the dermis) melanoma [5],[7]. About 6,850 patients suffering fatally from melanoma are likely to have died in 2020 [1]. Unfortunately over the past 40 years, melanoma cases have been steadily rising [2]. Amidst this melonomic gloom, the good news is that melanoma can be cured through excisions when detected and diagnosed in its early stages [14],[12]. At present, the available detection and diagnosis options for melanoma are visual inspection, clinical screening, dermoscopic analysis, biopsy and histopathological examination of skin lesion. Among all options, dermoscopy is the most popular imaging technique. Dermoscopy refers to microscopic examination and evaluation of skin lesions. It is typically done with every high quality magnifying lens and powerful illumination system (aka Dermatoscope [4]). However, dermoscopic images are not easy to interpret for

*Advised by: Mr. Jeremy Irvin, Stanford University.

diagnosis. Even with most experienced dermatologists, the evaluation of dermoscopic images can be laborious and error prone [13], [8]. The complex visual characteristics of skin lesions such as multi-sizes, multi-shapes, fuzzy boundaries, and low contrast when compared to the skin and noise presence such as skin hair, oils, air, and bubbles limit even an expert dermatologists's sensitivity to less than 80% [25]. Figure 1. gives a visual feel of some lesion images as classified by multiple expert dermatologists into benign and malignant melanoma.

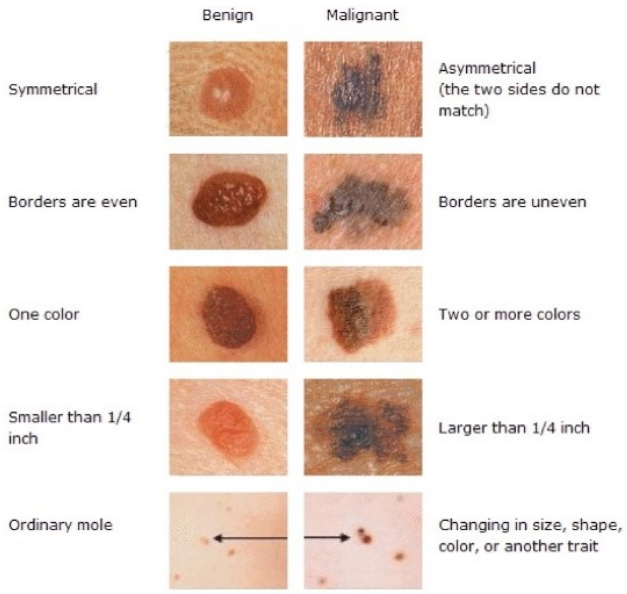

Figure 1: Two sets of dermoscopic images. Images on the left column show benign cases, while the images on the right show malignant cases

Aforesaid challenges especially motivate the machine learning community to design algorithms to automatically diagnose melanoma in dermoscopic images. Computer-aided diagnosis (CAD) system automates interpretation of dermoscopic images to diagnose melanoma. This helps in early and successful diagnosis of melanoma, thereby making the treatment effective and reducing the mortality rate. Machine learning methods aim to 'train' models using labeled data (dermoscopic image of benign and malignant melanoma) and then provide prognosis on the new dermoscopic images of patients. Recently, a subfield of machine learning known as deep learning has shown success in automatic medical image interpretation comparable to the level of human specialists. Deep learning focuses on the construction of "neural networks", which comprises multiple layers of non-linear functions with coefficients/weights which are derived from the training data. These weights are the result of the 'training' process and hold the knowledge of differentiation between benign and malignant melanoma. Artificial Neural Networks (ANNs) were the first type of neural network to

demonstrate success in medical imaging classification, including CadE/CadX to diagnose chest diseases given radiographs [23], diagnose general cancer given tumor and lymphatic node images [22], and diagnose prostate cancer given MRI images [16]. Researchers have developed specialized neural networks known as convolutional neural (CNN) networks which are designed to model the highly structured data present in images. Convolutional neural networks have been known for being effectively utilized to diagnose cancers and diseases in the past [15]. Some recent examples of using CNNs for medical-imaging related tasks include diabetic retinopathy detection which performed on levels comparable to ophthalmologists [18], chest X-Ray pathology detection in chest radiographs which matched the performance of radiologists [10] , and knee abnormality detection in MRI scans [11]. Given the recent success of convolutional neural networks on a variety of medical imaging tasks, there is a significant opportunity for research on developing models for other tasks [26]. One of the early works using CNN to classify dermoscopic images into benign and malignant, with accuracy levels comparable to that of 21 board-certified dermatologists, came from Esteval et. al in 2017 [17]. In this research paper, I developed deep learning models to classify dermoscopic images of skin lesions. The networks were trained and validated on a large dataset of dermoscopic images that were labeled by dermatologists as benign or malignant. I discovered that a ResNet model with 50 layers achieves an AUROC score of 0.78 on the validation set, while an EfficientNet model achieves a AUROC score of 0.90. The model leverages augmentations and ensembling of multiple smaller models to achieve that high performance on the validation set. Finally, I interpreted the model predictions through the use of Class Activation mappings to highlight where in the picture the model considered a possibly malignant tumor existed.

2 Methods

2.1 Data

The dataset consisted of 33,126 dermoscopic images in total and was split into a training set (25,932 images) to learn model parameters and a validation set (7,194 images) to compare models. As shown in Table 1., the training set consisted of 473 positive (malignant) cases or 1.8 % of the training set. The 25,459 images remaining were negative (benign) cases or 98.2 % of the training set. The validation set consists of 7,194 images. As shown in Table 1, the validation set consisted of 111 positive (malignant) cases or 1.6 % of the validation set. The 7,083 images remaining were negative (benign) cases or 98.4 % of the validation set. Both datasets had the exact same approximate age of 45, with the percent of female patients varying from 48.9% in the training set to 45.9% in the validation set. All images in the training and validation are in JPEG format and were resized to 224 x 224 pixels.

	Training	Validation
Positive No (%)	473 (1.8%)	111 (1.6%)
Negative, No (%)	25,459 (98.2%)	7,083 (98.4%)
Mean age	45.0	45.0
% Female	48.9 %	45.9 %
Total	25,932	7,194

Table 1: Data statistics across the data splits.

2.2 Convolutional Neural Networks

Convolutional neural networks are especially effective in machine learning due to their ability to leverage the structured format of imagery. Convolutional networks are composed of convolutional layers, which use a kernel and stride to extract certain features from the image. Conv layers are followed by non-linear Rectifier layers (aka ReLU in Figure 2.) which typically remaps the input to a manageable 'range' (e.g. -1 to +1). Typical CNNs are combinations of the three layers (convolution, rectifier, pooling). The final layer of a CNN is a fully-connected layer to produce the final output of the network (classification label). Certain CNN architectures are more effective for different types of tasks. To mitigate overfitting, two candidate models were investigated in this work for their relatively small neural network size - ResNet50 [19] and EfficientNet [24]. A diagram of the ResNet50 architecture is shown below with Figure 2. and the EfficientNet Architecture is shown in Figure 3.

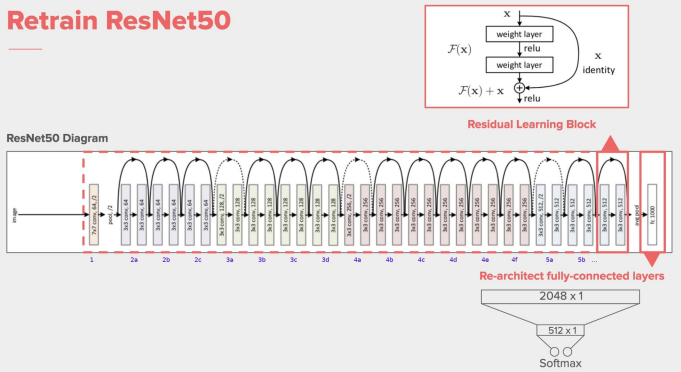

Figure 2: Diagram of the ResNet50 Architecture

In essence, the ResNet50 architecture functions by stacking up sets of the three layers 50 times [19]. The triplet consists of convolutional layers, normalization / pooling layers, and rectifier layers. The convolutional layer extracts

relationships from the image, the pooling layer aggregates the extracted features, the rectifier nonlinearities (ReLU) are applied to capture nonlinear structures within the image. One aspect that differentiates ResNet from other CNNs - is that it trains to learn the 'residual signal', which refers to the difference of a triplet's output with input of the previous triplet.

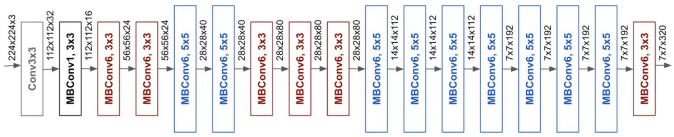

Figure 3: Diagram of EfficientNet Architecture Used

The EfficientNet architecture is composed of layers similar to the ResNet50 architecture but also efficiently balances the tradeoff between network depth (number of layers), width (number of channels), and image resolution. It is 8.4 times smaller and 6.1 times faster than the best convolutional neural network and achieves one of the highest accuracies on the ImageNet database, a large dataset of natural images which is the most commonly used benchmark for assessing convolutional neural network performance 24.

2.3 Training Procedure

A special training procedure was created to maximize the performance of the ResNet50 / EfficientNet models. In order to utilize the advantages of transfer learning, both models had weights initialized from a network pre-trained on data from ImageNet. Both models used a scheduler that reduced the learning rate each time performance on the validation set plateaued, an Adam optimizer, and the binary cross entropy loss function. The Adam optimizer was utilized with an initial learning rate of 0.001. The scheduler had a patience of 1, a threshold of $1 * 10^{-}5$ to measure when the validation performance plateaued, and a factor of 0.2 to reduce the learning rate by. Each model was trained for a maximum of 20 epochs (20 full passes over the training set) with a batch size of 32 and 4 workers (processors). The following augmentations were applied to each batch during training: random horizontal and vertical flips with a probability of 50%, random crops of the image, and random rotations of the image. After the model completed iteration over the training dataset, the model was directly tested on the validation set. The summary metric AUROC was computed for each epoch and the 5 models with the highest AUROC score were saved. These five models were used to generate an ensemble of models where the average of the probabilities predicted by each model was used as the prediction for the ensemble, as depicted in Figure 4. Finally, to limit training time and prevent overfitting on the training set, early stopping was used which terminates training once the marginal gains from training were lower than a specific threshold.

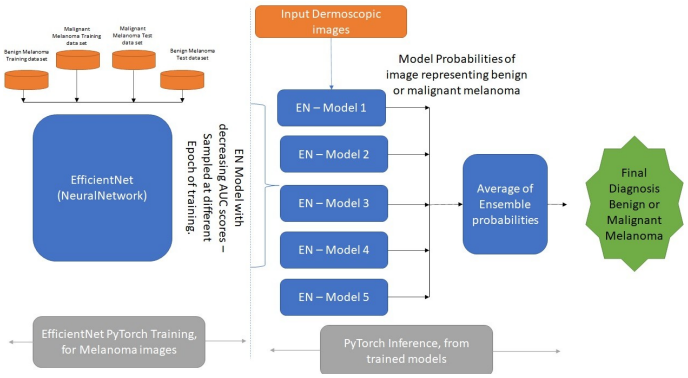

Figure 4: A Conceptual Diagram of how melanoma diagnosis is accomplished in the model.

2.4 Evaluation Metrics

I used a variety of performance metrics to evaluate the deep learning models. To evaluate the quality of the model probabilities, I computed the receiver operating characteristic (ROC) curve and precision recall curve (PR) . Then, to compress the curves into more comprehensive metrics, the area from the ROC and PR curves was computed to get AUROC and AUPRC. The model was primarily evaluated using the AUROC score, but AUPRC was also measured to provide a holistic evaluation of the models. To convert probabilities to binary predictions and compute point metrics, I used the threshold which led to the highest F1 score on the validation set. After the probabilities were converted to binary predictions at the optimal threshold, I computed the precision, which measures how many accurate malignant diagnoses were made over all malignant diagnoses the model predicted, recall, which measures how many accurate malignant diagnosis were made over all test images which were malignant, F1 score which is the harmonic mean of precision and recall, and accuracy of the model.

2.5 Model Interpretation

Once training and testing was complete, I sought to interpret the model predictions. In order to accomplish this, I used class activation maps (CAMs, Figures 5/6.) to highlight the regions of the image which contribute most to the model's prediction of melanoma. The class activation maps are computed using the feature maps right before the global average pooling layer (GAP) is reached, where feature map (1 per channel) is multiplied by the weight corresponding to the final fully connected layer, and then summed to create the CAM. The CAM is a heatmap that overlays the image where temperature/color highlights the regions of the image which contribute most to the model's classification. In Figures 5/6, the blue regions highlight areas which contribute more to the prediction of melanoma, while purple highlights areas which contribute less. Us-

ing these CAMs, especially for false negative / positive diagnosis cases, helped me understand possible causes that were derailing the model from an accurate prediction.

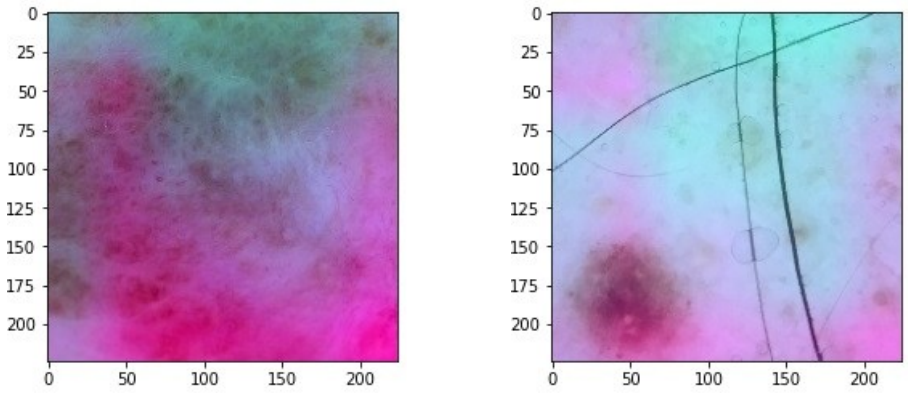

Figure 5: Class activation maps (CAM) of the best model on the validation set. The CAM on the left are an accurate diagnosis of benign melanoma, while the CAM on the right shows a false diagnosis (the model incorrectly predicted there was melanoma). Areas that have higher hues of blue indicate the model has a higher confidence that melanoma exists in that part of the image, while areas that are more purple indicate lower confidence.

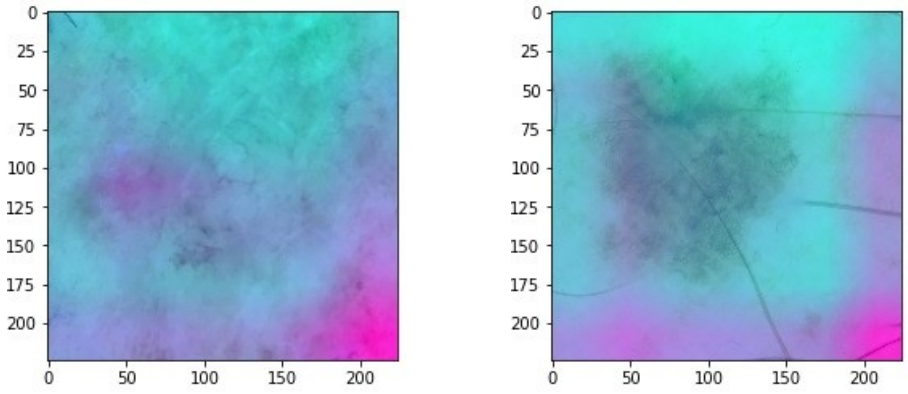

Figure 6: Class activation maps (CAM) of the best model on the validation set. The CAM on the left are an accurate diagnosis of malignant melanoma, while the CAM on the right shows a false diagnosis (the model incorrectly predicted there wasn't melanoma). Areas that have higher hues of blue indicate the model has a higher confidence that melanoma exists in that part of the image, while areas that are more purple indicate lower confidence.

3 Results

		Point Metrics				Summary Metrics	
	Experiment	Precision	Recall	F1	Accuracy	AUROC	AUPRC
1	Pretrained ResNet50, with optimizer Adam, no early stopping, and CrossEntropyLoss loss function	0.03	0.91	0.06	0.56	0.79	0.05
2	Pretrained ResNet50, with optimizer Adam, no early stopping and BinaryCrossEntropyWithLogits (BCEL) loss function	0.13	0.32	0.19	0.96	0.86	0.15
3	Pretrained ResNet50 with optimizer Adam and scheduler, no early stopping, and BCEL loss	0.09	0.51	0.15	0.91	0.87	0.11
4	Pretrained ResNet50, with optimizer Adam and scheduler, early stopping, and BCEL loss	0.07	0.71	0.13	0.85	0.88	0.14
5	Pretrained EfficientNet with optimizer Adam and scheduler, early stopping, and BCEL loss	0.14	0.14	0.04	0.89	0.54	0.02

6	Pretrained EfficientNet with optimizer Adam and scheduler, early stopping, BCEL loss, and ensembling	0.21	0.20	0.20	0.98	0.87	0.13
7	Pretrained EfficientNet with optimizer Adam and scheduler, early stopping with increased threshold, BCEL loss and ensembling and data augmentation (Error in generating predictions)	0.42	0.16	0.23	0.98	0.57	0.08
8	Pretrained EfficientNet with optimizer Adam and scheduler, early stopping with highest threshold, BCEL loss and ensembling	0.1	0.66	0.28	0.91	0.90	0.17

Table 2: Performance metrics of the experiments on the validation set.

I experimented with a variety of models and training procedures in order to investigate their impact on performance, primarily in terms of AUROC.. Starting with a ResNet50 model and Experiment 1 metrics as a baseline, the loss function was changed to BinaryCrossEntropyWithLogits (BCEL) due to its specialty in binary classification, which led to an increase inAUROC score from 0.79 to 0.86 (Experiment 2). Next, a scheduler was added in order to anneal the learning rate (Experiment 3). Although the AUROC score does increase from Experiment 2 to 3, the gain is relatively small. In Experiment 4, early stopping was added to prevent overfitting on the training set which also led to minimal performance gains. In Experiment 5, the ResNet50 model being used

was replaced with an EfficientNet model. This caused the AUROC score to plummet. In Experiment 6, 5 EfficientNet models were put into an ensemble where the predictions would come as the arithmetic mean of the 5 predictions of the individual models. This increased AUROC from 0.54 to 0.87. The intention for Experiment 7 was to train the model more rigorously by augmenting the training set images. Once the bug was fixed, the model outputted the highest and final AUROC score at Experiment 8 - 0.9.

4 Discussion

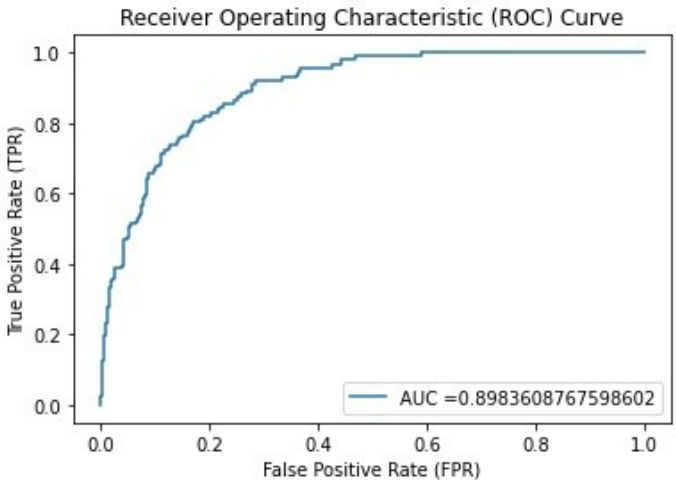

Figure 7: ROC curve of best model of the validation set.

In this work, I developed convolutional neural networks for the detection of melanoma in dermoscopic images. Two different models, ResNet50 and EfficientNet, were investigated together with a variety of training procedures. Through each experiment, one important variable was toggled, such as model type, loss function, use of scheduler, early stopping, data augmentation, and ensembling. Overall, the key methods used for the greatest gains in AUROC came from the use of early stopping and ensembling of the models. Apart from the model improvement methods, other techniques such as CAMs / graphs were utilized to figure out optimal values for certain variables and to analyze the models.

Major implications of our work lie mainly in the use of multiple techniques applied, and the gain of each technique. The set of techniques and their respective gains can be used by other researchers in the medical field to prioritize which techniques to use to maximize their own models' AUROC score. A second implication lies in the model's AUROC score. As the dermatologist accuracy of 75% [21] has been surpassed by our model's 91% accuracy and 90% ROC AUROC,

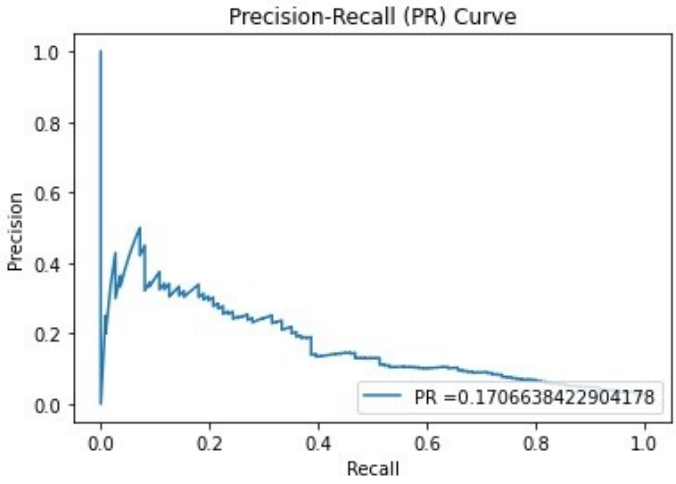

Figure 8: PR curve of best model of the validation set.

the model proposed has the potential to match or exceed human performance, implying it may be viable to assist humans in the detection of melanoma. This work also has important limitations which should be considered. First, data only came from 6 hospitals (Hospital Cl´ınic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, The University of Queensland, and the University of Athens Medical School), so the model may not be representative of certain populations [6]. As a result, the model will likely be less generalizable and over-represent the populations from where the dataset originated. Another limitation was the oversimplification of the task. In the case melanoma is classified, the severity must also be determined. One metric to determine this is called the Clark Scale, which ranks detected melanoma into 5 levels [20]. However in this work, melanoma was only classified based on existence, ignoring many nuanced details that are important to the diagnosis of cancer. Finally, skin lesions in different parts of the body may be treated differently. The models presented in this work do not utilize information about where the lesions originated in the body, which may be important information for classification. All such known limitations could be addressed with help of availability of a variety of data which covers different geographies, races, regions of the body, and severity levels of the melanoma. In future research, we plan to explore other model architectures, and versions of ResNet and EfficientNet which are available in PyTorch. Using K-Fold cross validation instead of saving the top 5 models would also be considered in order to improve the heterogeneity of models in the ensemble. Finally, generating custom augmentations such as inserting obstructions like hairs into the images for the training set could also improve the robustness of the model.

References

[1] American cancer society. https://cancerstatisticscenter.cancer.org/!/cancer-site/Melanoma Accessed: 2020-10-21.

[2] American cancer society. https://cancerstatisticscenter.cancer.org/cancer-site/Melanoma Accessed: 2020-10-21.

[3] Cancer. https://www.who.int/news-room/fact-sheets/detail/cancer. Accessed: 2020-10-21.

[4] Dermoscopy. https://dermnetnz.org/topics/dermoscopy/. Accessed: 2020-10-21.

[5] Melanoma statistics. https://www.curemelanoma.org/about-melanoma/melanoma-statistics-2/. Accessed: 2020-10-12.

[6] SIIM-ISIC melanoma classification. https://www.kaggle.com/c/siim-isic-melanoma-classification/data. Accessed: 2020-10-14.

[7] Skin cancer facts & statistics. https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/. Accessed: 2020-10-21.

[8] Qaisar Abbas, M E Celebi, Carmen Serrano, Irene Fondón García, and Guangzhi Ma. Pattern classification of dermoscopy images: A perceptually uniform model. *Pattern Recognit.*, 46(1):86–97, January 2013.

[9] Australian Institute of Health and Welfare. Cancer in australia: Actual incidence data from 1982 to 2013 and mortality data from 1982 to 2014 with projections to 2017. *Asia Pac. J. Clin. Oncol.*, 14(1):5–15, February 2018.

[10] Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for Multi-Label chest X-Ray classification. *Sci. Rep.*, 9(1):6381, April 2019.

[11] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, Safwan Halabi, Evan Zucker, Gary Fanton, Derek F Amanatullah, Christopher F Beaulieu, Geoffrey M Riley, Russell J Stewart, Francis G Blankenberg, David B Larson, Ricky H Jones, Curtis P Langlotz, Andrew Y Ng, and Matthew P Lungren. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Med.*, 15(11):e1002699, November 2018.

[12] Germán Capdehourat, Andrés Corez, Anabella Bazzano, Rodrigo Alonso, and Pablo Musé. Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions. *Pattern Recognit. Lett.*, 32(16):2187–2196, December 2011.

[13] M Emre Celebi, Hitoshi Iyatomi, William V Stoecker, Randy H Moss, Harold S Rabinovitz, Giuseppe Argenziano, and H Peter Soyer. Automatic detection of blue-white veil and related structures in dermoscopy images. *Comput. Med. Imaging Graph.*, 32(8):670–677, December 2008.

[14] M Emre Celebi, Hassan A Kingravi, Bakhtiyar Uddin, Hitoshi Iyatomi, Y Alp Aslandogan, William V Stoecker, and Randy H Moss. A methodological approach to the classification of dermoscopy images. *Comput. Med. Imaging Graph.*, 31(6):362–373, September 2007.

[15] S Charan, M J Khan, and K Khurshid. Breast cancer detection in mammograms using convolutional neural network. In *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, pages 1–5, March 2018.

[16] Pieter J L De Visschere, Alberto Briganti, Jurgen J Fütterer, Pirus Ghadjar, Hendrik Isbarn, Christophe Massard, Piet Ost, Prasanna Sooriakumaran, Cristian I Surcel, Massimo Valerio, Roderick C N van den Bergh, Guillaume Ploussard, Gianluca Giannarini, and Geert M Villeirs. Role of multiparametric magnetic resonance imaging in early detection of prostate cancer. *Insights Imaging*, 7(2):205–214, April 2016.

[17] 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(7639):115–118, February 2017.

[18] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C Nelson, Jessica L Mega, and Dale R Webster. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22):2402–2410, December 2016.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. December 2015.

[20] Breslow depth and clark level. https://www.curemelanoma.org/about-melanoma/melanoma-staging/breslow-depth-and-clark-level/. Accessed: 2020-10-14.

[21] A Naeem, M S Farooq, A Khelifi, and A Abid. Malignant melanoma classification using deep learning: Datasets, performance measurements, challenges and opportunities. *IEEE Access*, 8:110575–110597, 2020.

[22] Seyedmehdi Payabvash, Kaan Meric, and Zuzan Cayci. Differentiation of benign from malignant cervical lymph nodes in patients with head and neck cancer using PET/CT imaging, 2016.

[23] Chunli Qin, Demin Yao, Yonghong Shi, and Zhijian Song. Computer-aided detection in chest radiography based on artificial intelligence: a survey. *Biomed. Eng. Online*, 17(1):113, August 2018.

[24] Mingxing Tan and Quoc V Le. EfficientNet: Rethinking model scaling for convolutional neural networks. May 2019.

[25] M E Vestergaard, P Macaskill, P E Holt, and S W Menzies. Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting, 2008.

[26] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Others. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.*, 155(10):1135–1141, 2019.