Deep Learning Model Performances for Diagnosis of Skin Diseases

Sauman Das[1] and Spandan Das[1]

[1] *Thomas Jefferson High School for Science and Technology*

Abstract

A study conducted by Center for Disease Control (2016) estimates the occurance of 82,476 Melanoma cases every year. Melanoma is one of many prevalent classes of skin cancer. Providing a quick and non-invasive form of early diagnosis can greatly reduce the risk of fatality. Deep learning architectures can efficiently provide an accurate diagnosis given a highly magnified image of the skin. This study compares the performances of two Convolutional Neural Network (CNN) models, DenseNet121 and InceptionResNetV2, each trained with 20,265 images to predict 8 different skin diseases taken from the International Skin Imaging Collaboration (ISIC) database. The Area Under the Receiver Operating Characteristic (AUROC) curve score was used as a performance metric for evaluation. A 2-way ANOVA with replication was run to determine if the model and disease had a significant impact on the AUROC scores at a statistically significant level of $\alpha = 0.05$. The results show that the interaction between the model and the disease had a significant impact on the performance. Furthermore, a Tukey HSD Post-Hoc analysis shows that the InceptionResNet model performs best when predicting Basal Cell Carcinoma and Benign Keratosis. In this study, we conclude that the InceptionResNet model is the optimal architecture for diagnosing skin diseases.

1 Introduction

Deep learning has revolutionized the medical industry in recent years, particularly in disease diagnosis. The amount of medical imaging data has been increasing at unprecedented rates (De Fauw et al., 2018). This provides scientists with a massive amount of data which has spurred the adaptation of machine learning in medical science.

Currently, cancer is the second most fatal disease in the world. Skin cancer is known to be an aggressive form of cancer if not diagnosed at its early stages (Younis et al., 2019). To conquer this disease, Harvard released a data set called Human Against Machine 10000 in 2018 with around 10,000 dermatoscopic images with the corresponding diseases. In 2019, ISIC added more data to this collection consisting of 8 corresponding diseases (Codella et al., 2019).

- Melanoma
- Melanocytic Nevus
- Basal Cell Carcinoma
- Actinic Keratosis
- Benign Keratosis
- Dermatofibroma
- Vascular Lesion
- Squamous Cell Carcinoma

This study analyzes the differences in performance of an InceptionResNet and a DenseNet model, two state-of-the-art deep learning models, to gain a better understanding of the advantages of each of them with respect to medical computer vision tasks. The basic ResNet model was the winner of the ILSVRC competition in 2015 hosted by ImageNet (Srinivasan et al., n.d.). InceptionResNet was built on top of ResNet with the aim of achieving a tradeoff between performance and computational cost (Szegedy et al., 2017). On the other hand, DenseNets are more efficient because they are more compact networks. Both of these were trained with the same data. Thus, we hypothesize that the DenseNet and the InceptionResNet models would have a similar performance across all 8 test diseases.

Presence of heavy data imbalance in the data set posed a major challenge while training. One could achieve a 50.8% accuracy by simply predicting Melanocytic Nevus every time. Therefore, basic accuracy is not a good metric to use when evaluating this dataset. Instead, we use the AUROC score which takes into account the true and false positive rate for each disease. Previous studies have shown that this metric is more representative of the models' overall performance as opposed to simple accuracy measurements (Weng & Poon, 2008). Finally, a two-way ANOVA with post-hoc tests were run to validate the hypothesis.

2 Methods

2.1 Dataset

The ISIC 2019 training dataset consisted of 25,331 images collected from various locations on the body. There was an imbalanced distribution of the data among the 8 diseases which we addressed in preprocessing (Figure 1). For internal evaluation, we randomly split the data into training and validation sets with an 80-20 percent split (i.e. 20,763 images in the training set and 5,066 images in the validation set).

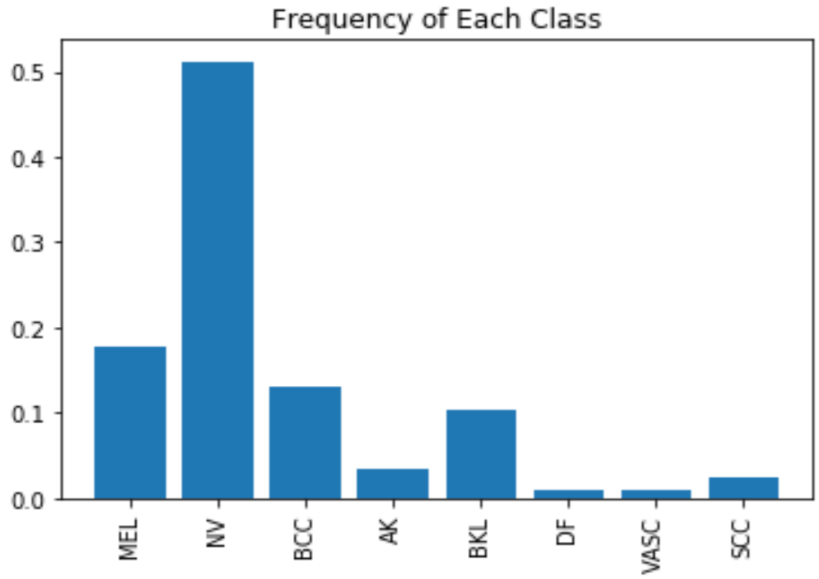

Figure 1: ISIC 2019 training set data distribution. Diseases include Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), and Squamous Cell Carcinoma (SCC).

2.2 Preprocessing

To prepare the images for training, we implemented three preprocessing steps. First, using batch statistics (as opposed to the entire dataset statistics), we normalized the mean and standard deviation of the training data. Moreover, we used the training data statistics to standardize the validation data. Next, we specified that the images were to be shuffled after each epoch of training, with the randomization acting as a form of regularization and reducing variance. Finally, we scale each image to a size of 224 by 224 pixels, since our models require a constant input dimensionality.

2.3 Training

To deal with the aforementioned class imbalances, our model (f) used a weighted variant of cross-entropy loss $\mathcal{L}$ defined for each training case x (with corresponding output y) as

$$\mathcal{L}(x) = -(w_p \cdot y \log(f(x)) + w_n \cdot (1-y) \log(1-f(x)))$$

where w_p and w_n are defined for each class as the frequencies of negative and positive instances, respectively (Coursera, 2020).

We trained both multi-class classifiers (DenseNet and InceptionResNet) by initializing their weights as those pre-trained on ImageNet. In order to adapt the models to our specific dataset, we added a 2-dimensional Global Spatial Average Pooling layer and a dense logistic output layer on top of the pretrained models. The models are trained using the Adam optimizer with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We trained both models for 5 epochs with a batch-size of 32. Training was performed using the Kaggle's standard NVIDIA Tesla P100 GPU.

3 Statistical Analysis

3.1 Data Collection

After the models were trained, each model was validated with 5 distinct datasets. The same preprocessing was applied to the validation sets as described in 2.2. Each model was run 5 times per disease for a total of 40 runs per model. Both models were run with identical validation sets (Figure 2).

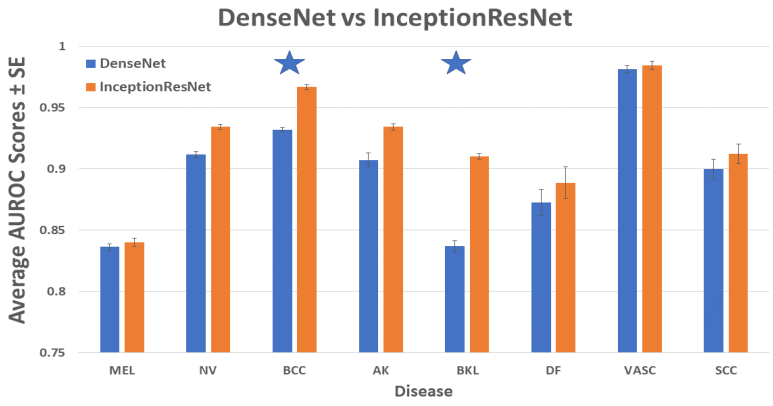

Figure 2: Average AUROC values (n=5) ± Standard Error (SE) for each Sample (stars represent statistical significance at $P < 0.05$ based on Tukey HSD pairwise comparisons). Diseases include Melanoma (MEL), Melanocytic Nevus (NV), Basal Cell Carcinoma (BCC), Actinic Keratosis (AK), Benign Keratosis (BKL), Dermatofibroma (DF), Vascular Lesion (VASC), and Squamous Cell Carcinoma (SCC).

Previously, we hypothesized that the model choice would not result in any difference in performances for any of the diseases. In general, both models perform very similarly, except for Melanocytic Nevus, Basal Cell Carcinoma, Actinic Keratosis, and Benign Keratosis (Figure 2).

3.2 ANOVA

A 2-Way ANOVA (n=5) was run to test the effect of both the model and disease on the AUROC score. The samples were collected from distinct sets and each group consisted of 5,066 images. To ensure that the data was sampled from an approximately normal distribution, the Shapiro-Wilks test was conducted using Python's SciPy stats package. Finally, the ANOVA was run with the collected data (Table 1).

Table 1: ANOVA Results ($\alpha = 0.05$)

Source of Variation	SS	df	MS	F	p-value	F crit
Disease	0.145	7	0.021	129.546	$P < 0.0001$	2.156
Model	0.012	1	0.017	72.752	$P < 0.0001$	3.991
Interaction	0.009	7	0.001	7.937	$P < 0.0001$	2.156

Based on the two-way ANOVA, a significant interaction ($P < 0.05$) suggests that both the model and the disease together have a significant impact on the AUROC score. In our hypothesis, we stated that the model choice would not result in a significant difference for any of the diseases. However, the interaction term tells us that the model did have a significant effect on the AUROC score for at least one disease.

3.3 Post-Hoc Tests

The ANOVA test demostrated that a significant difference existed, however, pairwise comparisons were needed to identify which scores varied significantly between models and disease. We ran a Tukey Honest Significance Difference (HSD) test using Python's *statsmodel* API to compare all model-disease pairs (Table 3).

Table 2: Results for the Post Hoc test show statistical significance

Tukey HSD Post Hoc Test ($\alpha = 0.05$)			
DenseNet Model	InceptionResNet Model	p-value	reject
Dense MEL	Resnet MEL	0.9	False
Dense NV	Resnet NV	0.2825	False
Dense BCC	Resnet BCC	0.005	True
Dense AK	Resnet AK	0.0831	False
Dense BKL	Resnet BKL	0.001	True
Dense DF	Resnet DF	0.7751	False
Dense VASC	Resnet VASC	0.9	False
Dense SCC	Resnet SCC	0.9	False

The results of this test show a statistically significant difference in the average scores of Basal Cell Carcinoma and Benign Keratosis which are represented

with the stars in Figure 2. The above results demonstrate that the IncpetionResNet model was the higher performing model. Tukey HSD results demonstrate that the InceptionResNet model made the most significant improvement when predicting Benign Keratosis.

4 Conclusion

In this paper, deep learning models for skin disease detection were investigated on the ISIC 2019 dataset. We introduced and studied two CNN models, useful for classifying images especially in a medical context. After both models were trained, they were evaluated based on their AUROC scores to determine the better performing model. We hypothesized that there would be no difference in model performance, regardless of the disease it was tested on. However, our data suggests that the InceptionResNet model performed superior to that of the DenseNet when predicting specific diseases such as Basal Cell Carcinoma and Benign Keratosis. Thus we believe that the InceptionResNet model should be the optimal choice for predicting skin diseases in general. In the field of medicine, the safest approach tends to be the best approach especially when diagnosing diseases. This study provides strong evidence to support that an InceptionResNet should be used as opposed to a DenseNet when predicting skin disease. The framework of this experiment paves the way for future research to study models and their impacts on diagnosing diseases.

References

Center for Disease Control. (2016). *USCS data visualizations - CDC.* `https://gis.cdc.gov/Cancer/USCS/DataViz.html`. Centers for Disease Control and Prevention. (Accessed 5/28/20)

Codella, N., Rotemberg, V., Tschandl, P., Celebi, M. E., Dusza, S., Gutman, D., ... others (2019). Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*.

Coursera. (2020). *AI for medical diagnosis.* `https://www.coursera.org/learn/ai-for-medical-diagnosis/home/welcome`. (Accessed 5/23/20)

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., ... others (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, *24*(9), 1342–1350.

Srinivasan, V., Zhang, Y., & Rezaee, M. (n.d.). Landuse/landcover classification using resnet50 in search for better test patch size.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first aaai conference on artificial intelligence.*

Weng, C. G., & Poon, J. (2008). A new evaluation measure for imbalanced datasets. In *Proceedings of the 7th australasian data mining conference-volume 87* (pp. 27–32).

Younis, H., Bhatti, M. H., & Azeem, M. (2019). Classification of skin cancer dermoscopy images using transfer learning. In *2019 15th international conference on emerging technologies (icet)* (pp. 1–4).