Effects of Clinical Data in Skin Cancer Classification

Lynn Kong(ldkong), Monica Pan(jpan5), Thomas Young(tomyoung)

1 Introduction

Skin cancer is the most common cancer in the United States. It is estimated that one out of five Americans will develop skin cancer by age 70. Unlike other types of cancer, skin cancer can be visually detected easily by dermatologists and it is highly curable if detected early. Published deep learning models [1, 2] for skin cancer classification demonstrates a viable diagnostic tool by skin images alone. Inspired by this publication, our project will build and analyze a baseline SVM classifier and a modified convolutional neural network (CNN) based on Google Inception V3 [3] to diagnose skin images as benign or malignant.

The input to each of our classifiers is a publicly sourced photographic image of skin. The input is processed through a model of either SVM or CNN with a final output of predicted clinical diagnosis for the image as benign or malignant.

We also further explored the effect of embedding clinical metadata in CNN training and evaluation. We generated multiple skin image data sets with different clinical metadata embedded in the fine-grained class labels, then trained and evaluated the CNN with each data set and inferred the final coarse-grain benign/malignant diagnosis from predicted classes. We observed that different clinical metadata has different effect on CNN performance in terms of accuracy, precision, and recall. Future exploration of compounding effects from different clinical metadata combination may lead to better CNN performance.

2 Related Work

Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep CNNs [4] show potential for general and highly variable tasks across many fine-grained object categories [5]. Esteva et. al. demonstrated classification of skin lesions using a single CNN [1, 2], trained end-to-end from images directly, using only pixels and disease labels as inputs. Esteva also outlined the development of a CNN that matches the performance of dermatologists at three key diagnostic tasks: melanoma classification, melanoma classification using dermoscopy and carcinoma classification.

Our original proposal was to completely replicate, analyze, and improve on Esteva's work. However, the published model is trained on multiple data sets that are not available to the public. More importantly, these data sets have over 2000 diagnosis labels that was key to the training class and inference algorithms in the publication. Thus, we chose to use the fine-grain disease to coarse-grain classification from the publication as inspiration and proceeded with our own fine-grain clinical data to coarse-grain classification model.

3 Dataset and Features

Images in data sets are from International Skin Imaging Collaboration (ISIC) Archive[6]. Each image in ISIC has a set of metadata, which may include the diagnosis (e.g. benign, malignant), anatomic site (e.g. lower extremity), age, and sex. We sourced 3028 images with the desired clinical metadata, with the total having roughly 3:1 ratio of benign to malignant images. From these images, we generated 6 data sets, D, DT, DS, DA, DTA, and $DTAS$. Each data set has 1876 training samples and 1152 test samples. The data sets differ in the label format of each sample, as shown in Figure 1. Data set D contains only diagnosis in sample label; DT contains diagnosis and anatomic site; DA contains diagnosis and age range; DS contains diagnosis and sex; DTA contains diagnosis, anatomic site, and age range; and $DTAS$ contains all for clinical metadata.

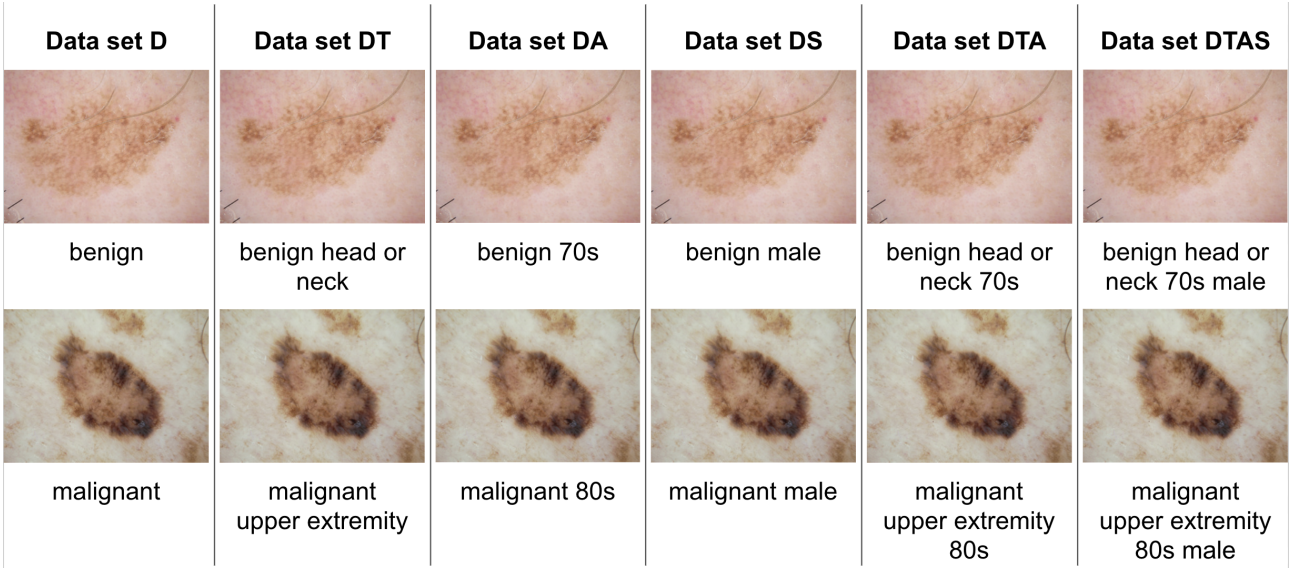

Figure 1: An example of how two images are labeled in the six data sets.

The source image dimension ratio were variable, so we set all samples in our data set to have a height of 299 pixels, with variable width. For SVM, each sample was cropped width-wise to a square image, and resized to 64×64 for faster training. For Inception v3, each sample was cropped width-wise to 299×299, per requirement of the Inception v3 architecture.

4 Method

4.1 SVM

We used support vector machines as our baseline, which is solved using the optimization problem $\min_{w,b} \frac{1}{2}\|w\|^2$ $s.t. \quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, \quad i = 1, \ldots, n$, as noted in lecture notes.
In our preliminary experiments, we looked at SVM with 4 different kernel functions:

- Linear: $K(x, y) = x^\top y + c$
- Radial basis function (RBF): $K\left(x, y\right) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$
- Sigmoid: $K(x, y) = \tanh\left(\alpha x^T y + c\right)$
- Polynomial: $K(x, y) = \left(x^\top y + c\right)^d$

We trained the SVM by benign/malignant on image data set D, during which we used 5-fold grid-search cross-validation to perform parameter optimization to obtain the best results.

4.2 CNN

The foundation of our CNN is the Google Inception V3 architecture. This established architecture contains 48 layers, employs batch normalization and dropouts during training, and calculates loss using a Softmax function, as shown in lecture notes:

$$\ell(\theta) = \sum_{i=1}^{n} \log \prod_{l=1}^{k} \left(\frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right)^{1\left\{y^{(i)}=l\right\}}$$

In our project, we used the Inception V3 CNN pre-trained on ImageNet [3] and removed the final classification layer. We then added our own classification layer depending on the labelling of the fine-grained data set. We

then infer the final coarse-grain classification of benign by summing the probabilities of all fine-grain training classes related to benign by the following equation,

$$P_b = \sum_{c \in V_b} P_c$$

where b means benign. Our model is sketched in Figure 2.

CNNs were trained and evaluated on an AWS GPU instance.

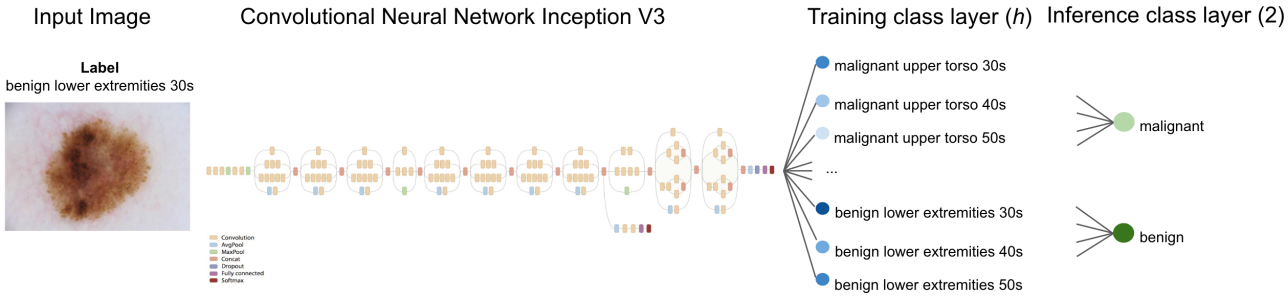

Figure 2: Proposed Inception v3 CNN Model with modified classification layers.

5 Experiments, Results, and Discussion

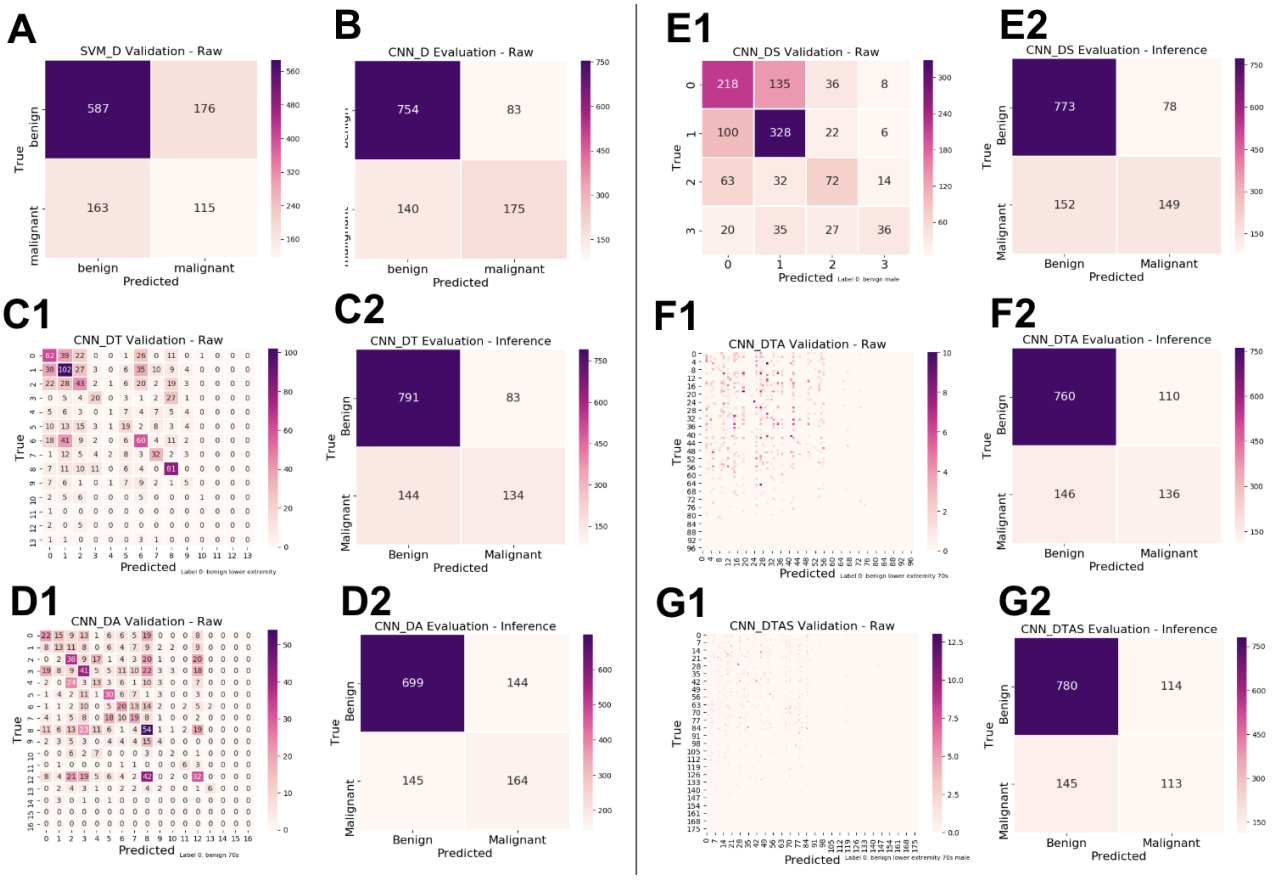

Figure 3: **A.** Confusion matrix of SVM trained and evaluated on data set D. **B.** Confusion matrix of CNN trained and evaluated on data set D. **C1-G2** Confusion matrices of CNN trained and evaluated on data sets DT, DA, DS, DTA, $DTAS$, with raw and inferred evaluation results.

5.1 SVM

SVM Among the SVM models, the one with the polynomial kernel function outputs the best result in terms of low false negative rate. Although the models with sigmoid and RBF kernels had the highest accuracy (not shown), the correct predictions solely came from the benign skin lesions image. Those two models did not identify any malignant skin lesion that could possibly lead to skin cancer, which is a big failure. We would like to avoid this situation in our model since a model that cannot correctly classify a potential malignant skin lesion fails to serve the purpose.

The test results of SVM with polynomial kernel had an accuracy is 67.44%, precision of 39.52%, recall of 41.37% F1 score of 0.4. We plotted the confusion matrix in Figure 3A.

5.2 CNN Training and Evaluation

From hereon, we will refer to a CNN trained on data set X as CNN_X.

During training, we tuned the threshold for the inference classification such that we achieve 91.5% specificity for all CNN. We then evaluated each model on their corresponding test set and plotted the confusion matrices of the fine-grained classification and the coarse-grained classification, as shown in Figure 3B-G2. We also determined the accuracy, precision, recall, and F1 score after the inference step for coarse-grain benign/malignant classification, as shown in Table 1.

Model	Inference Accuracy	Inference Precision	Inference Recall	F1 Score
CNN_D	0.8064	0.6783	0.5556	0.6108
CNN_{DT}	0.8030	0.6175	0.4820	0.5414
CNN_{DA}	0.7491	0.5325	0.5307	0.5316
CNN_{DS}	0.8003	0.6564	0.4950	0.5644
CNN_{DTA}	0.7778	0.5528	0.4823	0.5151
CNN_{DTAS}	0.7752	0.4978	0.4380	0.4660

Table 1: Performance of CNN trained on all data sets, tuned to achieve 91.5% specificity. Performance is measured after inference.

5.3 Discussion

We first observe that all CNN performed better than the SVM in terms of accuracy, precision, recall, and F1 score. This is expected, given that CNN is able to handle complex image classification better than SVM. Furthermore, this suite of CNNs should perform better in image classification in general because they contain the transfer learning from ImageNet.

Next, we look at the effects of the different fine-grain data sets on the CNN performance. At 91.5% specificity, we first see that CNN_D, trained with the most coarse-grained data set, performed best in terms of accuracy, precision, recall, and F1 score. Given our relatively small data set size, the fact that CNNs trained with more complex classification perform worse is expected. However, we do observe interesting comparisons between metrics of the CNNs trained and evaluated on clinical metadata labeling. We see that in comparison to CNN_{DT} and CNN_{DS}, CNN_{DA} performed the worse in terms of accuracy and precision. However, CNN_{DA} does have a higher recall (sensitivity), which is valuable to explore because higher sensitivity is desirable in diagnostic models. We also observe that the accuracy of CNN_{DTA} and CNN_{DTAS} falls between that of CNN_{DA} and CNN_{DT} or CNN_{DS}. The precision and recall of CNN_{DTA} also falls between that of CNN_{DT} or CNN_{DS}. The precision and recall of CNN_{DTAS} is the lowest out of all CNN. These results suggest that there is some significant compounding effect when different clinical metadata are combined in the fine-grain classification labeling, the exact nature of which needs to be explored further.

We need to do the following to correct certain errors/missing information in our analysis and make our project

more robust:

- Correct our hyper-parameter tuning method by separating a validation data set from the training set and tune based on results from validation data set.
- Train and evaluate CNN without ImageNet pre-training on all data sets. These model performances will serve as another set of baselines that will give us better context on the native behavior of CNN on skin image classification with different clinical metatdata.
- Tune the SVM baseline threshold such that it achieves 91.5% specificity, and evaluate and compare to CNN performance. This will provide a more reasonable comparison between the performances of the different types of models.
- Train and evaluate SVM on the 5 data sets with additional clinical metadata in the fine-grain labels. This will also provide a more reasonable comparison between the performances of the different types of models.

6 Conclusion and Future Work

We observed that CNN performs better than our SVM baseline in terms of accuracy, which is expected. We also observed that additional clinical metadata in the data set does result in changes in model performance. However, none of the CNNs trained on more fine-grained data sets performed better than the CNN trained on the data set with the binary benign/malignant labels, which may be due to the small data set sample size and the imbalanced fine-grain data sets.

Besides the immediate analysis follow-ups listed in the discussion, we also plan to do more work on the models and dataset generation. We plan to tune batch size as a hyperparameter for CNN. We also plan to train with a higher specificity target and and tune hyperparameters under those conditions. We can add more raw images to our data set to balance the data set and reduce the chance of confounding variable effect. We can also augment the data sets via translation, rotation and flip of existing images to introduce noise to data and reduce overfitting. Lastly, we plan to perform similar analysis on more clinical metadata, e.g. geographical region, in isolation and in combination with the metadata we have already analyzed, to find the best model performance in terms of accuracy, precision, and recall.

7 Source Code

Code and generated data are on Github [7].

8 Contribution

Lynn Kong worked on the data set generation from ISIC API and Inception V3 training and testing, analysis, as well as writeup. Monica Pan worked on the SVM baseline implementation, analysis, and writeup. Thomas Young worked on analysis and writeup.

References

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

[2] Yunzhu Li, Andre Esteva, Brett Kuprel, Rob Novoa, Justin Ko, & Sebastian Thrun, Skin Cancer Detection and Tracking using Data Synthesis and Deep Learning, arXiv:1612.01074, December, 2016

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, and Jonathon Shlens, Rethinking the Inception Architecture for Computer Vision, arXiv:1512.00567, December, 2015

[4] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature 521, 2015.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, Deep residual learning for image recognition, https://arxiv.org/abs/1512.03385, 2015.

[6] Dataset: International Skin Imaging Collaboration Archive, https://www.isic-archive.com

[7] https://github.com/ldezhenkong/cs229-project