Effects of Clinical Data in Skin Cancer Classification

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

CS229: Machine Learning
Autumn 2019

Abstract

Skin cancer can be visually detected and is highly curable if detected early. Inspired by a 2016 skin image classification model [1], we built a baseline SVM classifier and a CNN based on Google Inception V3 to diagnose skin images as benign or malignant. We further explored the effect of clinical metadata in CNN. We generated multiple skin image data sets with different clinical metadata embedded in the class labels, then trained and evaluated the CNN with each data set and inferred the final benign/malignant diagnosis from predicted classes. We observed that different clinical metadata has different effect on CNN performance in terms of accuracy, specificity, and sensitivity. Future exploration of compounding effects from different clinical metadata combination may lead to better CNN performance.

Data

Images in data sets are from ISIC [2]. Each image in ISIC has a set of metadata, including the diagnosis (e.g. benign, malignant), anatomic site (e.g. lower extremity), age, and sex. We sourced 3028 images total (30% in validation) with 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 1987 training samples and 1041 validation samples. The data sets differ in the label format of each sample. 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.

Models

SVM For our baseline model SVM, we experimented with four different kernel functions: linear, radial basis function (RBF), sigmoid and polynomial to classify by benign/malignant on image data set D. Then we used grid-search CV to perform parameter optimization to obtain the best results.

CNN We retrained Inception V3 CNN pretrained on ImageNet [3] with our generated data sets. We then infer the final classification of benign by summing the probabilities of all 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 the figure 1A. We trained and evaluated the modified CNN on all 6 data sets.

Source code Github repo [4].

Discussion

As expected, CNN greatly improved the binary classification results in comparison to baseline SVM, with all CNNs performing at higher overall prediction accuracy, specificity, sensitivity, and F1 score. Notably, the detective rate of malignant skin cancer from 0.41 in SVM to 0.56 in CNN_{DA}.

Out of the three clinical metadata added to sample label, age range has the most significant effect on model performance, as seen in metrics of CNN_{DA}, CNN_{DTA} and CNN_{DTAS} in Table 1. Models trained with age range in sample label has higher sensitivity, but with the drawback of lower accuracy and specificity. Higher sensitivity is especially desirable in clinical diagnostics, so this effect should be further explored. The drawback may be caused by the number of additional classes added by decade-based age range. Converting to age range with larger bucket size (e.g. 20 years or 30 years) may improve accuracy and specificity.

Future Research

We observed that each analyzed clinical metadata has a noticeable effect on CNN performance, whether in isolation and in combination with each other. First, we will resample the data with uniform distribution across labels to confirm that the effects due to confounding variables. Next, we plan to tune the age range metadata by increasing the bucket size of each age range. We also plan to perform similar analysis on more clinical metadata, e.g. geographical region. Lastly, we plan to find the best combination of clinical data in sample labels to increase model inference accuracy, specificity, and sensitivity.

Results

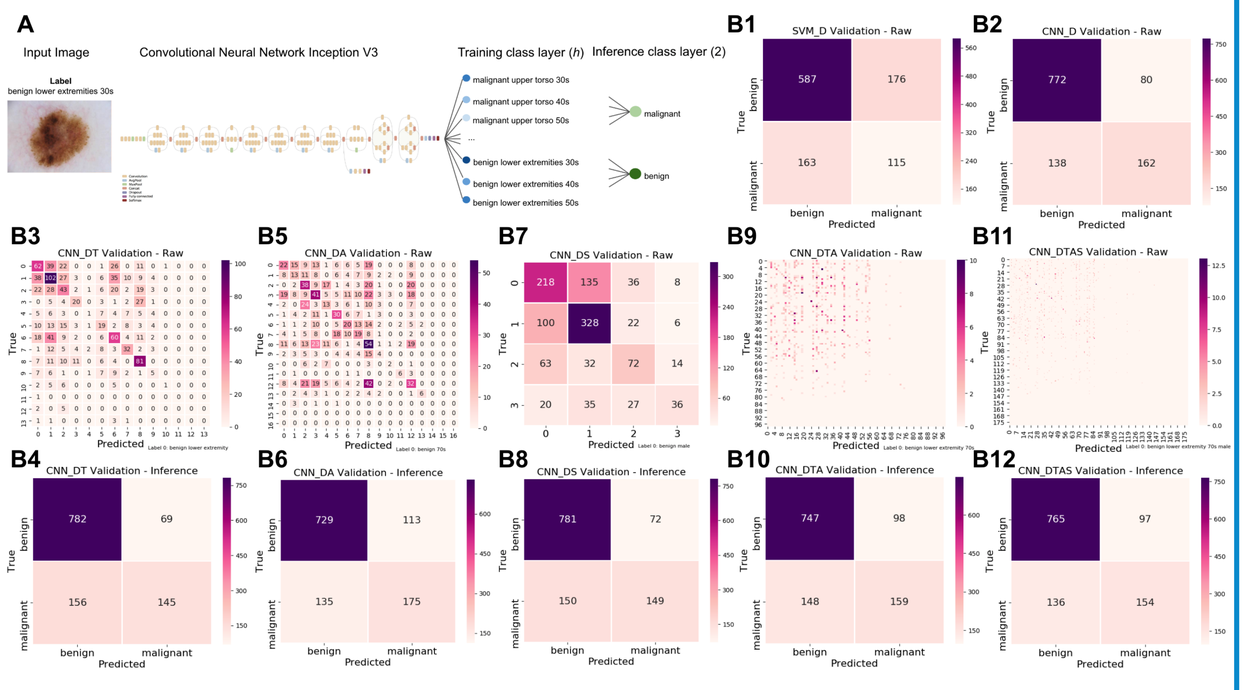

Figure 1: A. Modified CNN architecture, based on Inception V3. **B1.** Confusion matrix of SVM trained and evaluated on data set D. **B2.** Confusion matrix of CNN trained and evaluated on data set D. **B3-12.** Confusion matrices of CNN trained and evaluated on data sets DT, DA, DS, DTA, $DTAS$, with raw and inferred evaluation results.

Model	Inference Accuracy	Inference Specificity	Inference Sensitivity	F1 Score
SVM_D	0.6744	0.7693	0.4137	0.4042
CNN_D	0.8108	0.9061	0.5400	0.5978
CNN_{DT}	0.8047	0.9189	0.4817	0.5631
CNN_{DA}	0.7847	0.8658	0.5645	0.5852
CNN_{DS}	0.8073	0.9156	0.4983	0.5731
CNN_{DTA}	0.7865	0.884	0.5179	0.5638
CNN_{DTAS}	0.7977	0.8875	0.5310	0.5693

Table 1: Performance of all models, including SVM with polynomial kernel function trained on D and CNN trained on all data sets. Performance is measured after inference.

SVM Among the SVM models, the one with the polynomial kernel function outputs the best result in terms of low false negative rate. The resulting overall accuracy is 67% with a F1 score of 0.4.

CNN CNN_{DA} has lower inference accuracy and specificity, but higher sensitivity compared to model trained with CNN_D. CNN_{DS} and CNN_{DT} has similar inference accuracy, higher specificity, and lower sensitivity compared to model trained with CNN_D. CNN_{DTA} and CNN_{DTAS} has similar metric comparison as CNN_{DA}. CNN_{DTAS} had the highest F1 score.

References

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

[2] Dataset source: International skin imaging collaboration archive.

[3] Sergey Ioffe Christian Szegedy, Vincent Vanhoucke and Jonathon Shlens. Rethinking the inception architecture for computer vision. *arXiv*, 1512.00567, 2015.

[4] Source code. *GitHub*, https://github.com/ldezhenkong/cs229-project.