# Final Report for 6.862 project

Shaoling Han | shaoling@mit.edu

- **Introduction**

  Dermatosis is a major health burden for people nowadays. Early detection, distinguishing and diagnosis of skin cancers and benign dermatosis are of high importance. The main method of dermatosis diagnosis is based on observation of skin. This project focuses on applying machine learning methods to help classification of skin disease based on medical skin images and potentially other medical information.

- **Related work**

  There are some previous studies on this issue, mainly by using various architectures of CNN to classify the images. Philipp Tschandl *et al.* collect 139 machine learning algorithms and compare their accuracy against human readers, and find that machine learning algorithms are significantly better than human readers which is promising in future clinical practice(1). Amirreza Rezvantalab *et.al.* select 4 popular architectures (DenseNet 201, ResNet 152, Inception v3, InceptionResNet v2) and compare their performance on the combination of HAM10000 and PH2 dataset(2). Research from Aryan Mobiny *et.al.* also combine CNN with Bayesian methods. But these studies use only the image data(3). In this project, I'll try to build some simple CNN models based on previous research and explore potential approaches to utilize medical information other than images.

- **Data**

  The dataset is the HAM10000 dataset ("Human Against Machine with 10000 training images"), which contains 10015 dermatoscopic images (600×450, about 270KB in size) and information of these images, including patients' ID, image ID, ground truth disease diagnosis within one of 7 skin cancer/carcinoid categories (the label of images), method of obtaining ground truth, age, gender and localization of dermatosis(4). There are 9987 patients involved, so some patients contribute more than one images to this dataset.

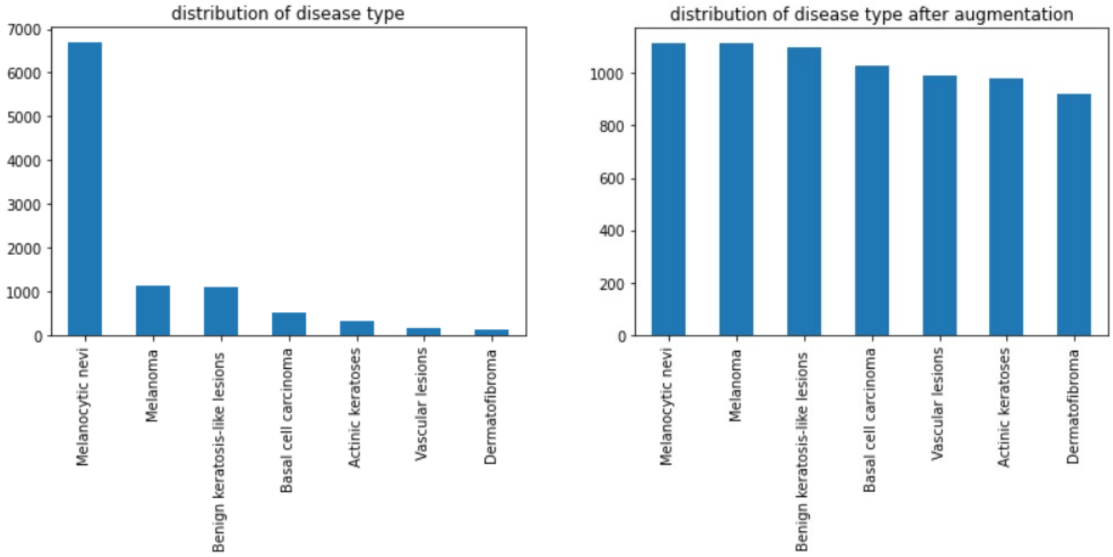


*Figure 1 Skin disease distribution, before augmentation (left) and after augmentation (right)*

- **Preprocessing**

  57 images are with missing age information. These NAs are simply replaced with the average of age. To fit in CNN models, the images are resized to 64×48/48×36/32×24 to keep the original width/length ratio, with both RGB and grayness channel. The values of pixels are divided by 255 to keep the values within 0 to 1. Given the imbalanced distribution among different diseases, data augmentation is implemented: each resized image is rotated with an angle randomly selected from (-30°, 30°) or (150°, 210°) with equal probability. To achieve

balanced distribution, 1117 images (about 1/6) are randomly selected among “melanocytic nevi”, and 2, 3, 7, 8 times many images are produced among “basal cell carcinoma”, “actinic keratoses”, “vascular lesions” and “dermatofibroma” respectively. See Figure 1 for details.

- **Methods**

  The main method used in this project is convolutional neural network (CNN). The input is numeric array representing image, and the output is predicted skin disease class. The loss function is cross entropy, a standard loss function used for classification problems. In this project I explore different variants of the main method: using input images with different size; using input images with RGB or grayness channel; using original or augmented images; using only images or images plus non-image information (age, sex, localization of skin disease) by concatenating flatten layer from CNN and non-image data. Please see the supplementary for detail architecture of CNN.

  The dataset is divided into training, validation and test partitions with ratio 6:2:2. There are several hyperparameters in CNN: learning rate, epochs number and batch size. The optimizer used in back propagation is “Adam”. I try several learning rate and find lr= 0.0001 gives acceptable results. I also use early stopping in training: if the accuracy in validation set doesn’t improve within 10~15 epochs, it stops training. In practice, all models end with early stopping within 100 epochs, so I set the max epochs as 100. I try several batch size and find that batch size can hardly change model performance, so I use a typical value 32. The models are trained with training data (using validation data for validation during training) and evaluated with test data (not touched before training). The main evaluation metric is overall classification accuracy. I also compute the area under curve (AUC) for each of the 7 categories to inspect the model’s distinguishability for different diseases.

- **Results**

  The performance of models is summarized in tables below. The max value for each column is marked red.

| Models (no augmentation) | | AUC | | | | | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| | | Actinic keratoses | Basal cell carcinoma | Benign keratosis-like lesions | Dermatofibroma | Melanocytic nevi | Melanoma | Vascular lesions | |
| 64×48 | RGB | 0.95 | 0.958 | 0.886 | 0.877 | 0.925 | 0.859 | 0.992 | 77.53% |
| | Gray | 0.932 | 0.907 | 0.809 | 0.82 | 0.862 | 0.772 | 0.689 | 71.54% |
| 48×36 | RGB | 0.943 | 0.943 | 0.892 | 0.893 | 0.917 | 0.837 | 0.984 | 76.14% |
| | Gray | 0.912 | 0.887 | 0.804 | 0.869 | 0.868 | 0.781 | 0.758 | 72.14% |
| 32×24 | RGB | 0.935 | 0.939 | 0.872 | 0.896 | 0.9 | 0.809 | 0.959 | 73.69% |
| | Gray | 0.903 | 0892 | 0.796 | 0.834 | 0.859 | 0.772 | 0.644 | 70.19% |

*Table 1 Model performance for images data without augmentation*

| Models (image augmentation) | | AUC | | | | | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| | | Actinic keratoses | Basal cell carcinoma | Benign keratosis-like lesions | Dermatofibroma | Melanocytic nevi | Melanoma | Vascular lesions | |
| 64×48 | RGB | 0.926 | 0.923 | 0.822 | 0.984 | 0.919 | 0.866 | 0.997 | 68.23% |
| | Gray | 0.874 | 0.842 | 0.707 | 0.966 | 0.827 | 0.789 | 0.934 | 53.76% |
| 48×36 | RGB | 0.921 | 0.922 | 0.821 | 0.979 | 0.91 | 0.879 | 0.995 | 66.23% |
| | Gray | 0.863 | 0.846 | 0.694 | 0.955 | 0.822 | 0.793 | 0.946 | 51.07% |
| 32×24 | RGB | 0.92 | 0.924 | 0.835 | 0.978 | 0.904 | 0.861 | 0.995 | 67.13% |
| | Gray | 0.854 | 0.803 | 0.704 | 0.934 | 0.817 | 0.803 | 0.921 | 50.45% |

*Table 2 Model performance for images data after augmentation*

| Models (image augmentation+non-image data) | | AUC | | | | | | | Accuracy |
|---|---|---|---|---|---|---|---|---|---|
| | | Actinic keratoses | Basal cell carcinoma | Benign keratosis-like lesions | Dermatofibroma | Melanocytic nevi | Melanoma | Vascular lesions | |
| 64×48 | RGB | 0.935 | 0.931 | 0.854 | 0.987 | 0.931 | 0.878 | 0.996 | 70.78% |
| | Gray | 0.872 | 0.851 | 0.719 | 0.968 | 0.859 | 0.79 | 0.945 | 53.34% |
| 48×36 | RGB | 0.921 | 0.923 | 0.849 | 0.987 | 0.926 | 0.872 | 0.997 | 69.26% |
| | Gray | 0.879 | 0.821 | 0.725 | 0.959 | 0.846 | 0.784 | 0.936 | 51.90% |
| 32×24 | RGB | 0.907 | 0.915 | 0.834 | 0.971 | 0.913 | 0.87 | 0.994 | 66.23% |
| | Gray | 0.854 | 0.833 | 0.72 | 0.969 | 0.84 | 0.784 | 0.949 | 54.24% |

*Table 3 Model performance for images data after augmentation, concatenated with age, sex and disease localization*

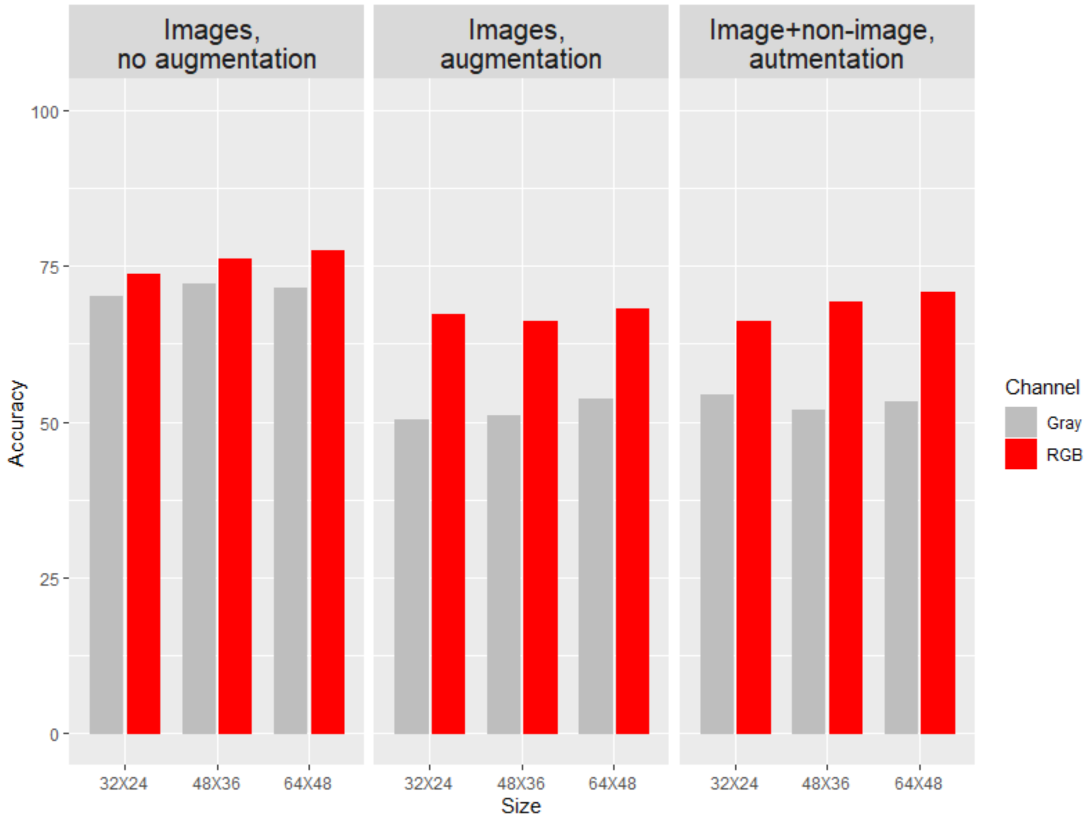


*Figure 2 Summary of models' performance*

From these results, the 64×48 RGB model outperforms others in all data preprocessing settings. However, the difference between using larger images and smaller images is minor. For example, when using augmented images, the 32×24 RGB model is slightly better than 48×36 RGB model; when using augmented images plus non-image data, the 32×24 Gray model is the best among all sizes.

The difference between RGB and Gray model is more obvious: when using images without augmentation, the RGB models are slightly better than corresponding Gray models; when using images with augmentation, the RGB models are significantly better than corresponding Gray models.

Augmentation indeed affect models' performance. Before augmentation, the accuracy is always higher than 70%, but after augmentation, the RGB models can merely reach 70% and the Gray models score below 55%.

Combining non-image data and images does not make significant difference to models' performance.

From the AUC results, benign keratosis-like lesions and melanoma seem more difficult to classify than other diseases, with or without image augmentation. After augmentation, dermatofibroma seems less difficult to classify, and vascular lesions seems less difficult to classify in Gray models (given that RGB models have already done pretty well even before augmentation).

- **Discussion**

Overall, the results indicate that having full color channels is more important than having larger image size, and image augmentation is necessary in such case of class imbalance.

It is reasonable that RGB images gives better prediction than Gray images: there is more information available in RGB images. Color of dermatosis area can certainly convey information about the disease type and progress. On the contrast, Gray images can hardly give color information, but only depicting the outline and shape of dermatosis area.

It is a bit surprising that image size is not playing a big role. I expected that images with higher resolution can convey more information thus yield better classification. The reason of the unexpected results could be: the CNN used here is not deep enough to capture more complicated latent features of the images, so larger image size doesn't help much; or it could be that diagnosis of dermatosis does not require super complex feature recognition so even 32×24 images are sufficient. More experiments can be done to test these hypotheses. If employing much more complicated CNN shows favor to larger images, then the current CNN is too shallow; if comparable performance can be achieved with simpler CNN, then dermatosis classification just does not require complicated feature extraction.

To illustrate the necessity of data augmentation, the confusion matrices of some models are plotted below.

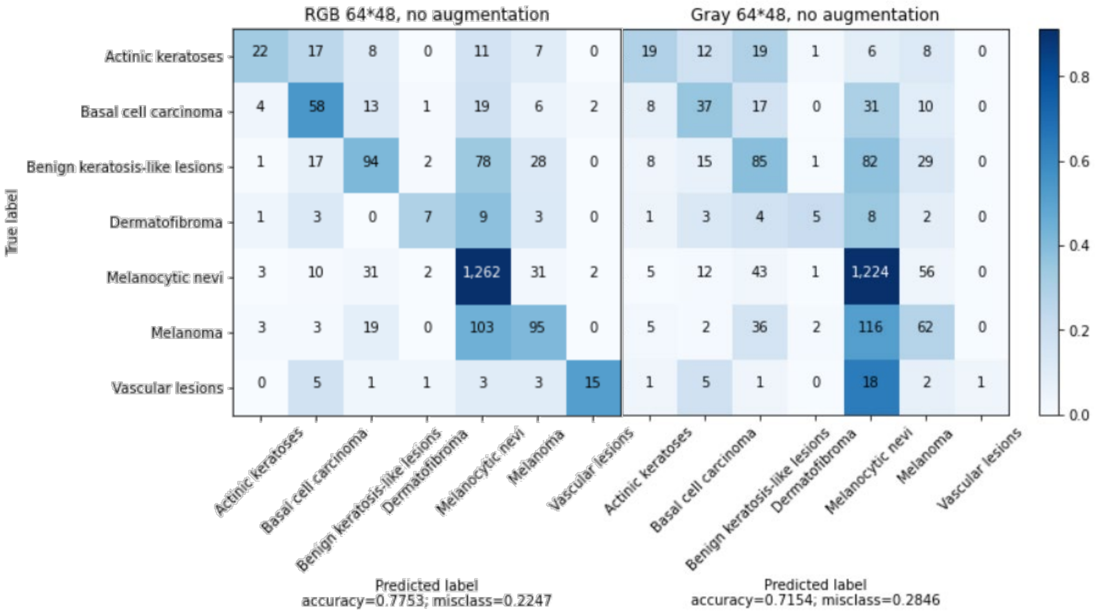
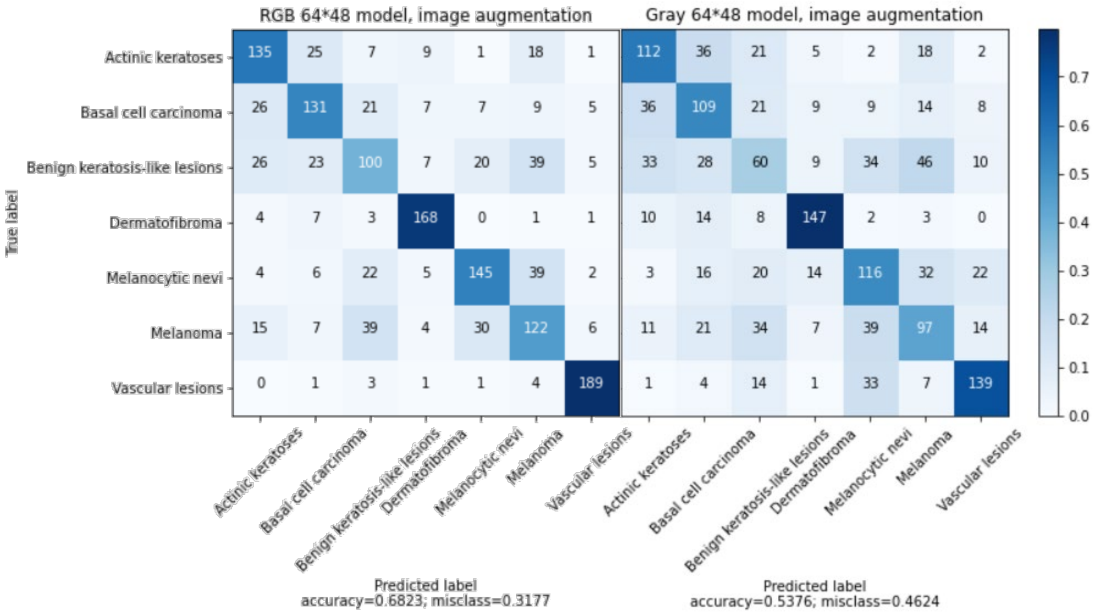


*Figure 3 Confusion matrix for no augmentation*

In the case of no augmentation, the predicted class is dominant in "melanocytic nevi", which has the most cases (much more than other diseases) in the whole dataset. For diseases other than “melanocytic nevi”, a lot of cases are still predicted as “melanocytic nevi”. For example, among 223 cases with ground truth “melanoma”, the RGB and Gray models predicted 103 and 95 of them as “melanocytic nevi”, respectively. In Gray model, it even only hit one correct among 28 cases of “vascular lesions”. Notice that even one blindly predicted every case as “melanocytic nevi”, there is still about 6705/10015=66.95% chance of correctness. So the high accuracy in image without augmentation set is biased and should be corrected.



RGB 64*48 model, image augmentation

| True label \ Predicted label | Actinic keratoses | Basal cell carcinoma | Benign keratosis-like lesions | Dermatofibroma | Melanocytic nevi | Melanoma | Vascular lesions |
|---|---|---|---|---|---|---|---|
| Actinic keratoses | 135 | 25 | 7 | 9 | 1 | 18 | 1 |
| Basal cell carcinoma | 26 | 131 | 21 | 7 | 7 | 9 | 5 |
| Benign keratosis-like lesions | 26 | 23 | 100 | 7 | 20 | 39 | 5 |
| Dermatofibroma | 4 | 7 | 3 | 168 | 0 | 1 | 1 |
| Melanocytic nevi | 4 | 6 | 22 | 5 | 145 | 39 | 2 |
| Melanoma | 15 | 7 | 39 | 4 | 30 | 122 | 6 |
| Vascular lesions | 0 | 1 | 3 | 1 | 1 | 4 | 189 |

Gray 64*48 model, image augmentation

| True label \ Predicted label | Actinic keratoses | Basal cell carcinoma | Benign keratosis-like lesions | Dermatofibroma | Melanocytic nevi | Melanoma | Vascular lesions |
|---|---|---|---|---|---|---|---|
| Actinic keratoses | 112 | 36 | 21 | 5 | 2 | 18 | 2 |
| Basal cell carcinoma | 36 | 109 | 21 | 9 | 9 | 14 | 8 |
| Benign keratosis-like lesions | 33 | 28 | 60 | 9 | 34 | 46 | 10 |
| Dermatofibroma | 10 | 14 | 8 | 147 | 2 | 3 | 0 |
| Melanocytic nevi | 3 | 16 | 20 | 14 | 116 | 32 | 22 |
| Melanoma | 11 | 21 | 34 | 7 | 39 | 97 | 14 |
| Vascular lesions | 1 | 4 | 14 | 1 | 33 | 7 | 139 |

*Figure 4 Confusion matrix for augmentation*

On the contrast, in the image with augmentations set, the models give more reasonable prediction for each disease. Although the RGB model has accuracy 68.23%, lower than Gray model without augmentation, it can handle all type of diseases properly. The diagonal of the matrix is more consistently dark, representing higher accuracy among each disease. Such model is more useful than the model trained without augmentation.

This can also partially explain the results of AUC. Dermatofibroma and vascular lesions are the two diseases with least cases n original dataset. After augmentation and balancing, the model is more adequately exposed to images of these diseases, and they are easier to recognized and predicted.

It is not surprising that adding non-image data doesn’t help much. These dermatoscopic images focus on minor part of human body, and they are not likely linked with macro aspects of the patients, such as age and sex. And the dimension of flatten CNN is way more than these non-image data, so concatenating them straight forward may not be a good idea.

- **Future work**

  Based on current results, there are some improvements that could be made in future. First, more image sizes and CNN architectures could be tested to see how important image size is. Second, during the augmentation many original images of melanocytic nevi are thrown away, which is a waste of data. Using weight loss function may deal with class imbalance and make the most advantage of data at the same time. Third, the model architecture still need refinement to link image and non-image data together. More dense layers with less bottleneck dimension after flatten CNN layer may improve the combination of these two types of data.

- **Reference**

1. Tschandl P, Codella N, Akay BN, Argenziano G, Braun RP, Cabo H, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for

pigmented skin lesion classification: an open, web-based, international, diagnostic study. The Lancet Oncology. 2019;20(7):938-47.

2. Amirreza Rezvantalab HS, Somayeh Karimijeshni. Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms 2018. Available from: https://arxiv.org/abs/1810.10348.

3. Mobiny A, Singh A, Van Nguyen H. Risk-Aware Machine Learning Classifier for Skin Lesion Diagnosis. Journal of clinical medicine. 2019;8(8).

4. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Available from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T.

- **Supplementary**

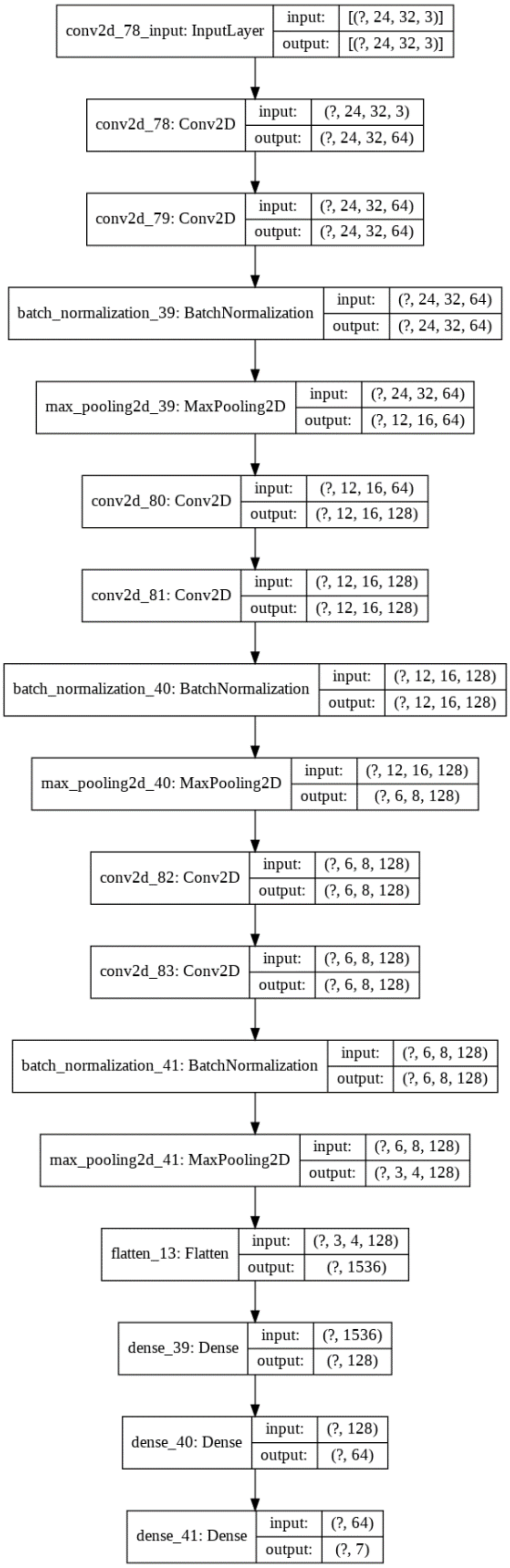

conv2d_78_input: InputLayer | input: [(?, 24, 32, 3)] | output: [(?, 24, 32, 3)]
conv2d_78: Conv2D | input: (?, 24, 32, 3) | output: (?, 24, 32, 64)
conv2d_79: Conv2D | input: (?, 24, 32, 64) | output: (?, 24, 32, 64)
batch_normalization_39: BatchNormalization | input: (?, 24, 32, 64) | output: (?, 24, 32, 64)
max_pooling2d_39: MaxPooling2D | input: (?, 24, 32, 64) | output: (?, 12, 16, 64)
conv2d_80: Conv2D | input: (?, 12, 16, 64) | output: (?, 12, 16, 128)
conv2d_81: Conv2D | input: (?, 12, 16, 128) | output: (?, 12, 16, 128)
batch_normalization_40: BatchNormalization | input: (?, 12, 16, 128) | output: (?, 12, 16, 128)
max_pooling2d_40: MaxPooling2D | input: (?, 12, 16, 128) | output: (?, 6, 8, 128)
conv2d_82: Conv2D | input: (?, 6, 8, 128) | output: (?, 6, 8, 128)
conv2d_83: Conv2D | input: (?, 6, 8, 128) | output: (?, 6, 8, 128)
batch_normalization_41: BatchNormalization | input: (?, 6, 8, 128) | output: (?, 6, 8, 128)
max_pooling2d_41: MaxPooling2D | input: (?, 6, 8, 128) | output: (?, 3, 4, 128)
flatten_13: Flatten | input: (?, 3, 4, 128) | output: (?, 1536)
dense_39: Dense | input: (?, 1536) | output: (?, 128)
dense_40: Dense | input: (?, 128) | output: (?, 64)
dense_41: Dense | input: (?, 64) | output: (?, 7)

*Figure S1 Example CNN architecture for image-only models. The image size is applied to change.*

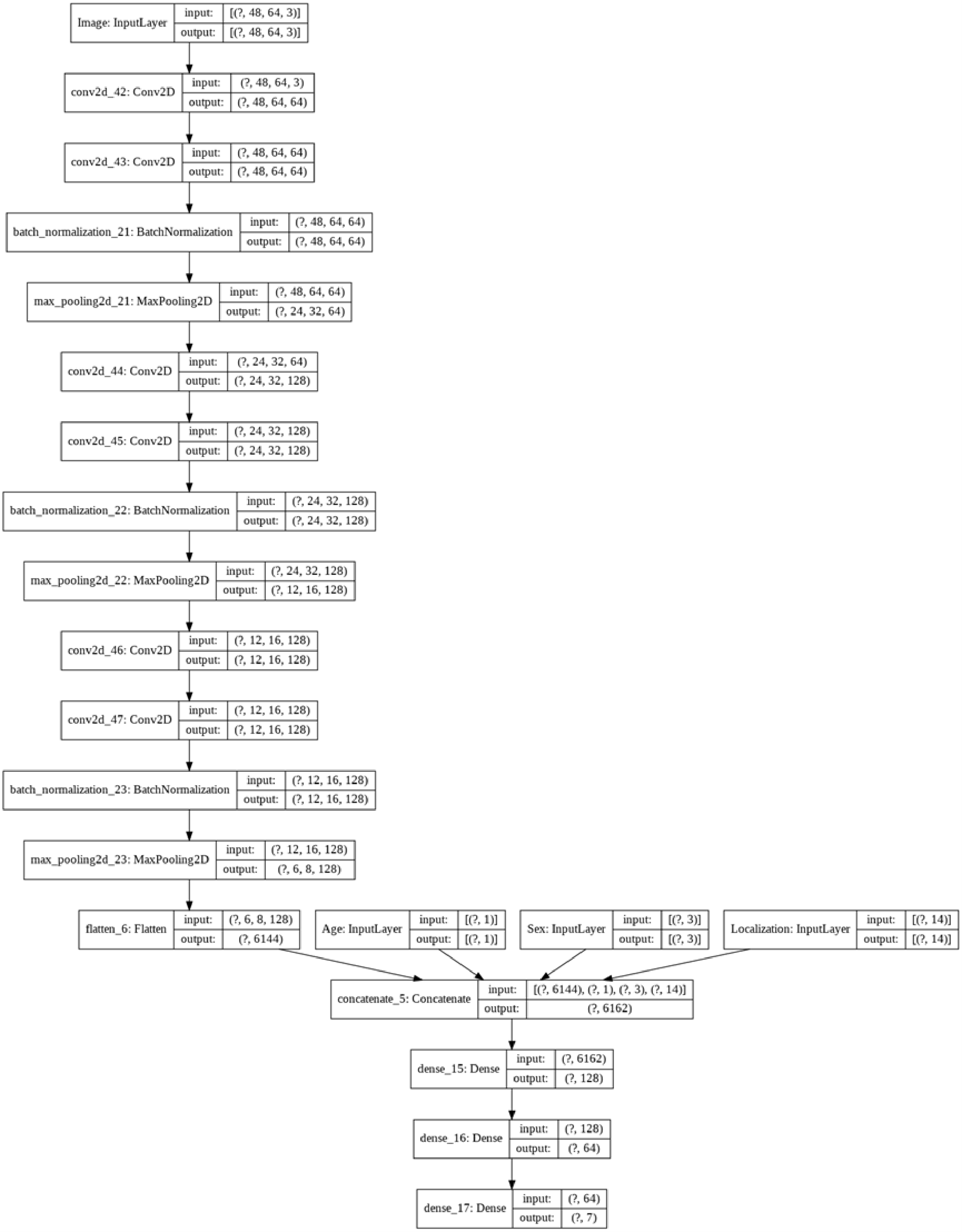


*Figure S2 Example CNN architecture for image+age+sex+disease localization models. The image size is applied to change.*