SIIM-ISIC Melanoma Classification

Abstract:

Identifying moles as benign or malignant class visualizing skin lesion images is a challenging task in dermatology. An efficient computer vision algorithm can avoid human error on visual inspection of lesion images saving time for further diagnosis. Using deep learning neural network and transfer learning architectures, machine learning binary models are developed to identify lesion images to benign or malignant. The models are built using publicly available data in Kaggle's competition SIIM-ISIC Melanoma Classification. CNN model built with four convolutional layers and two dense layers classifies lesion images with a precision of 0.48, recall of 0.92, F1 score of 0.63 and AUC of 0.75. The model developed with transfer learning approach using VGG19 and inceptionV3 performed better than CNN.

Introduction:

Early detection of skin cancer can be treated effectively with appropriate treatment. However, the visual inspection of medical images for skin cancer detection depends on the doctor's perception, and the task is tedious and time-consuming. An alternative approach for early detection is the efficient computer vision algorithm that can automate image classification with minimal error. A convolutional neural network (CNN) trained on thousands of medical images can extract features from new images and classified them into benign or malignant categories.

In medical images, malignant mole differs from benign in color, shape and size. Malignant signs include asymmetric shape, irregular border and appearance with varying colors such as red, pink, white, blue and black. These features are imprinted in the image with varying pixel values. Reading the pixel intensities, the CNN model can classifies skin lesions as benign or malignant.

CNN consists of several convolutional networks mixed with nonlinear and pooling layers and fully connected layers. The convolutional network captures spatial and temporal image features using multiple kernels/filters. While the first convolutional layer extracts low features such as edges, color, gradient, orientation, and corners by reading image pixels, the inner convolutional layer extracts high features such as corners and combinational edges using the previous layer. In convolution operation, convolved features obtained with the kernel are activated by the nonlinear layer and then feed into the pooling layer to reduce the spatial size of the convolved feature i.e, downsampling the image volume retaining the most important features needed to identify images. The CNN model can be built either by combining convolutional and neural network layers from scratch or using already trained models as a starting point through a transfer learning approach.

Data:

Society for Imaging Informatics in Medicine (SIIM) in collaboration with International Skin Imaging Collaboration (ISIC) has archived dermoscopic images of skin lesions. Data is available in Kaggle through a competition 'SIIM-ISIC Melanomic Classification 2020'. The training data consists of 33,126 images with metadata such as patients id, sex, age, anatomic site, diagnosis and lesion label as benign or malignant with number 0 and 1. Only around 1.76% (584) of images are malignant images implying the highly imbalanced dataset (Fig. 1). There are 2,056 patients and 21.82% of them have at least one malignant images. The number of images per patient ranges between 2 to 115. The metadata indicates lesion images from six anatomic sites such as torso, lower extremity, upper extremity, head/neck, palms/shoes and oral/genital (Fig. 2a). The number

of images from male and female sexes are 47.45% and 52.55%, respectively. Malignant lesion observed across all ages of patient ranging from 15 to 90 years old (Fig. 2b).

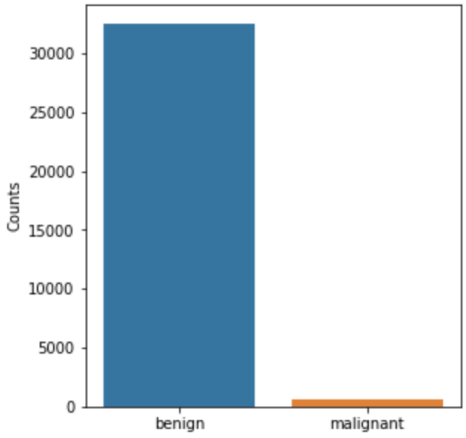

Figure 1: Bar diagram displaying number of benign and malignant images

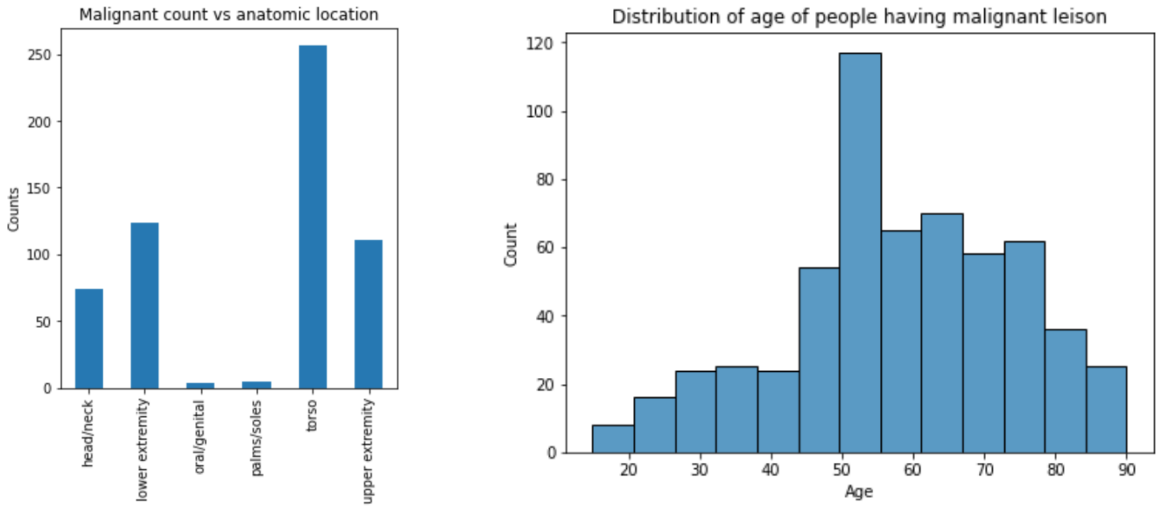

Figure 2: Bar diagram showing a) anatomic location of malignant images and b) age distribution of patients having malignant images.

Image preprocessing:

The quality of original images can be enhanced by image pre-processing steps such as image enhancement, image restoration and morphological methods. Image enhancement includes image scaling to reduce all images to same pixel size, color transformation (RGB to grayscale), contrast enhancement to improve the brightness difference between the foreground and background via histogram equalization or adaptive histogram equalization. Image restoration involves restoration of blurry and noisy images using smoothing methods such as Gaussian, Salt and Pepper, Poisson and Speckle smoothing or denoising methods such as spatial filtering and transform domain filtering.

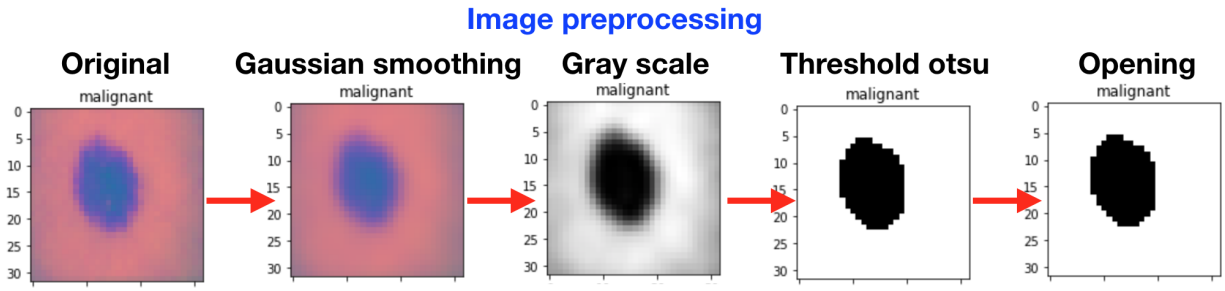

Figure 3: Image pre-processing. a) Original malignant image b) Gaussian smoothing c) color transformation d) Threshold image with otsu and e) Morphological erosion followed by dilation (opening).

For the binary images, morphological methods such as erosion, dilation or both can be applied to identify shape of lesion, edges and borders. While erosion reduces object area and features removing small white noises, dilation increases object area and accentuate features. Opening is the erosion followed by dilation and helps in increasing area removing the noises.

Data augmentation:

Image augmentation can be applied to expand the size of a training dataset by creating modified version of images. The dataset can be artificially expanded through different augmentation techniques such as random rotation, shift, shear, flip, feature standardization, brightness variation, blurring, etc. The increased dataset helps to develop a robust and generalized model dealing with a wide range of possible data. Image augmentation can be performed without much effort using the ImageDataGenerator class from the deep learning network Keras library.

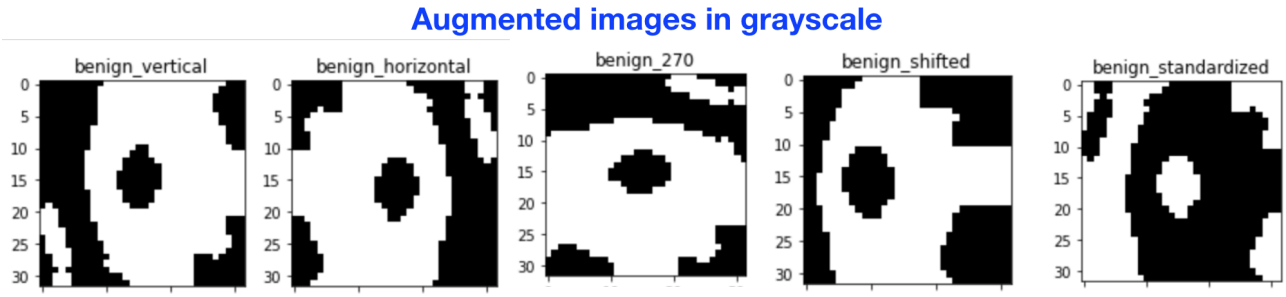

Figure 4: Augmented gray scale images with a) vertical flip b) horizontal flip c) rotation d) shift and e) feature standardization.

Model:

Image classification algorithms are built on equalized dataset consisting of equal number of randomly selected benign and malignant images. The equalized dataset is partitioned into train, validation and test sets with a size ratio of 0.8, 0.1 and 0.1. As the equalized dataset is smaller, data augmentation techniques are employed to expand the training data set during the model building. Images are normalized to the pixel values ranging between 0 and 1, and are trained using grayscale and RGB channels.

The CNN model is built using four convolutional layers and two neural network layers (Fig. 5). The four convolution layers are 1) 32-3x3, 2) 64-3x3, 3) 128-3x3 and 4)128-3x3 filters. Each convolution layer is mixed with nonlinear ReLu activation function and max-pool layers having size 2x2 filters with strides 2. After applying 4 convolution and max-pooling operations, the down-sampled outputs are fed into two fully connected neural network layer with 256 and 1 neurons, respectively. The final output is obtained with sigmoid activation function. The model is trained using Adam as optimizer, binary cross entropy as loss function and AUC as evaluation metrics. To prevent overfitting, regularization techniques such as dropout with 0.2 are used between neural network layers in combination with early stopping criteria and data augmentation. The model is trained both on colored and gray scale images having size 32 by 32 pixels.

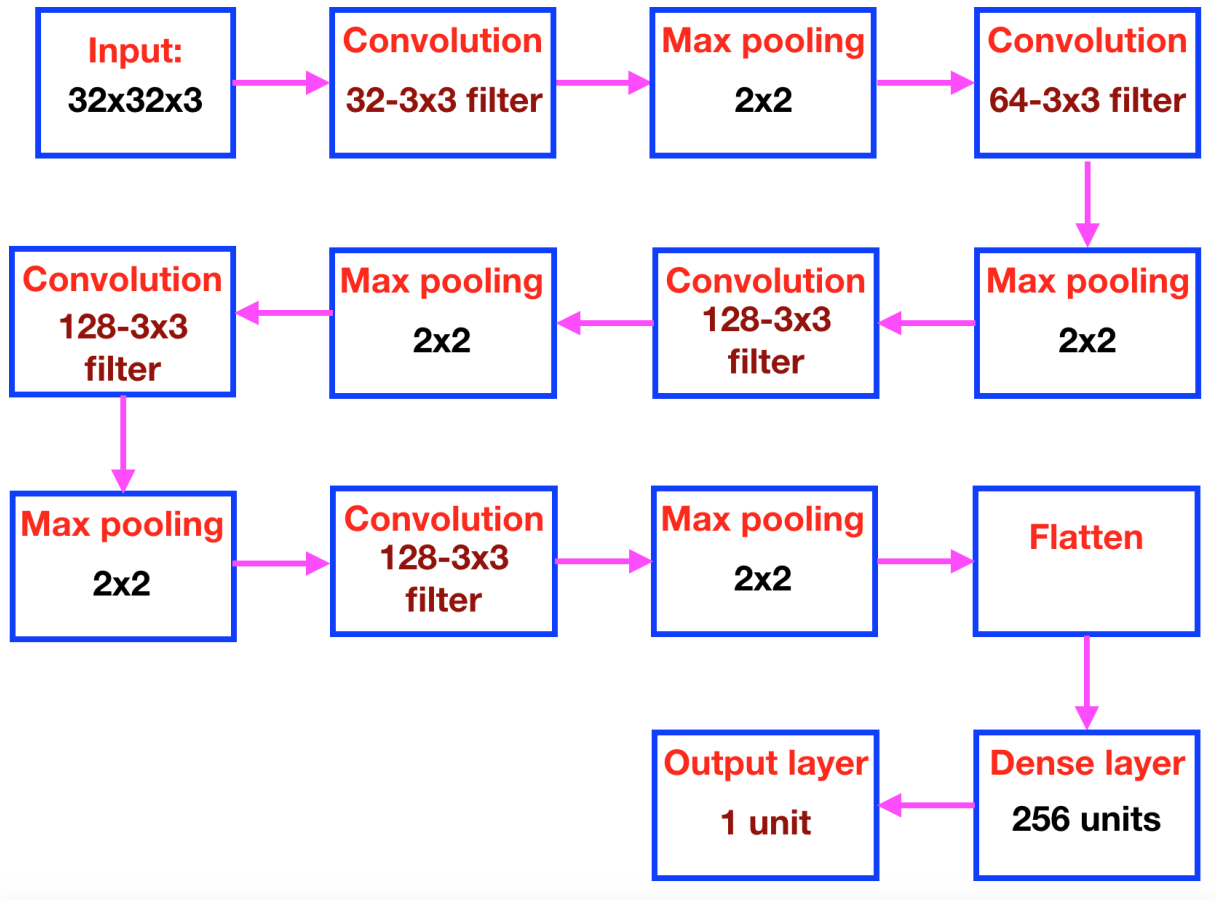

Figure 5: CNN architecture

Apart from CNN model, transformer models namely VGG19, Resnet50, Resnet152, InceptionV3, and EfficientNet B0 and B5 are trained freezing all layers or allowing few layers trainable.

Results:

The CNN model performance is evaluated passing the test set images. The model performance with colored images is observed to be better than gray scale images with morphological methods. Thus, the colored images are only used for fine tuning the CNN architecture and transformer models.

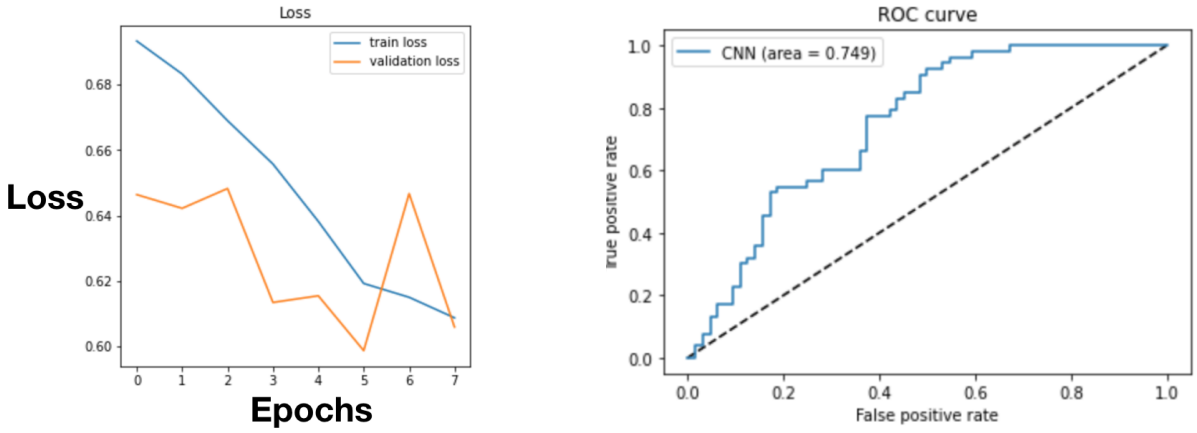

Figure 6: CNN model performance. a) Training and validation loss vs epoch b) ROC-AUC curve

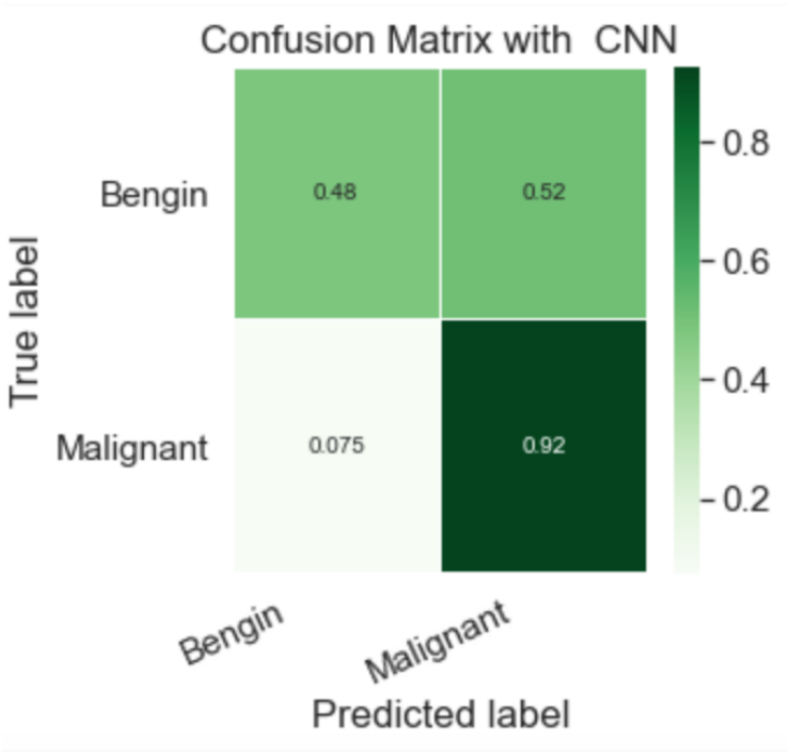

Figure 7: Confusion matrix

Figure 6a shows the training and validation loss as the CNN model continued training with a learning rate of 1e-3 and a batch size 4. After 8 epochs, the training stops with 0.59 training loss and AUC of 0.749 (Fig 6b). The evaluation scores of CNN model are 0.48 precision, 0.92 recall and 0.63 F1 score. Thus, the model is critical in predicting malignant images than benign images. As the cost of incorrectly classifying malignant images are much risker than incorrectly classifying benign images, the model reasonably fulfills the objective for identifying malignant images.

The performance of transfer based models are shown in Table 1 with different evaluation metrics. VGG19 and inceptionV3 model performed better than CNN on precision, F1 score, and AUC.

However, the Resnet and EfficientNet couldn't be performed better than CNN.

Table 1: Table showing model performance with different architectures.

Model	Loss	AUC	Precision	Recall	F1-score
CNN	0.59	0.75	0.48	0.92	0.63
VGG19	0.63	0.79	0.55	0.87	0.67
Resnet50 and 152	0.71	0.70	0	1	0
InceptionV3	0.60	0.80	0.55	0.89	0.68
EfficientNetB0 and B5	0.695	0.39	0	1	0

Recommendations:

One should expect a model with the best recall to avoid misclassification of malignant images and a better F1 score to improve the classification of benign images. The InceptionV3 model is better suited for classifying medical images into benign and malignant classes among the various models developed. Also, its recall is comparable to the CNN model. Thus, the hospital and physician could use it to detect early signs of malignant in medical images.

Future scopes:

All models are developed using equalized dataset. One can build model performance with the imbalanced dataset and fine-tune with hyperparameters such as batch size, learning rate, average pooling, filter size, and strides. Also, the model performance with other pre-trained models such as Xception, VGG16, MobileNet, Densenet, NASNet, MobileNetV2 can be evaluated. The pre-trained models can also be fine-tuned freezing all layers or making few layers trainable.

Conclusions:

Machine learning models with CNN architecture and transfer-based models are developed for skin lesion classification into benign and malignant classes. The models are developed utilizing data augmentation techniques on equalized datasets. The CNN model predicted the highest recall but with lower precision compared to transfer-based models. The inceptionV3 model is better suited for classifying medical images with higher recall and F1 scores than other developed transfer-based models.