FINAL CAPSTONE PROJECT REPORT

Skin Lesion Classification on the HAM10000 dataset using CNN

By: Daniel Logan

Introduction

This report summarizes the process of using deep learning to predict 1) whether a particular skin lesion has an elevated risk of cancer and 2) the classification of a skin lesion.

Practical Applications of Findings

Skin cancer is an ailment which affects millions of people every year. It is estimated that 1 in every 5 Americans will experience skin cancer at some point in their life, and it can pose a significant danger. However, even for the most dangerous type of skin cancer, Melanoma, the 5-year survival rate is 99 percent, meaning that early detection and treatment drastically improve patient outcomes.

Problem Statement

The problem for this capstone is using deep learning CNN models to accurately categorize skin lesions in order to improve early detection both by providing preliminary diagnoses as well as providing patients with care during periods between dermatology appointments. If our model were able to detect 90% of cases where lesions are problematic, it could save lives by informing these people of the need for an appointment and lessen strain on medical infrastructure in cases where patients do not need care.

Data Utilized

For this project, I utilized the HAM10000 dataset, a Harvard dataset of about 10,000 images of skin lesions within six of the most important categories for diagnosis of lesions, as well as some simple demographic information on the patients. This dataset was created in order to provide data for the training of CNN models which could compete in classification tasks against subject matter experts.

Pre-modeling Data Manipulation

One initial step for preprocessing taken was the resizing of our images to allow for further manipulations and running of more complex models while utilizing less processing power. Prior to utilizing our original dataset for both our binary classification problem and our categorical classification problem, we have significant issues regarding class imbalance in our original dataset which must be addressed. As we can see in Figure 1, some classes have significantly more data than others. Additionally, as can be seen in the color scheme of Figure 1, our binary category "Cancer Risk" (orange) has much more data than the other category, "Low Cancer Risk" (blue). In order to overcome this, I created a number of images which were copies of original images in non-dominant classes by flipping images horizontally, vertically, and both. Unfortunately, for our multi-class data, two categories had to be excluded, "Vascular Lesions" and "Dermatofibroma", due to insufficient data present in our dataset to have confidence in our model's generalizability for new lesions. These category drops, image flips, and some image cloning and image deletion, allowed for a perfect balancing of classes for our binary classification problem and near perfect balance for our categorical classification problem.

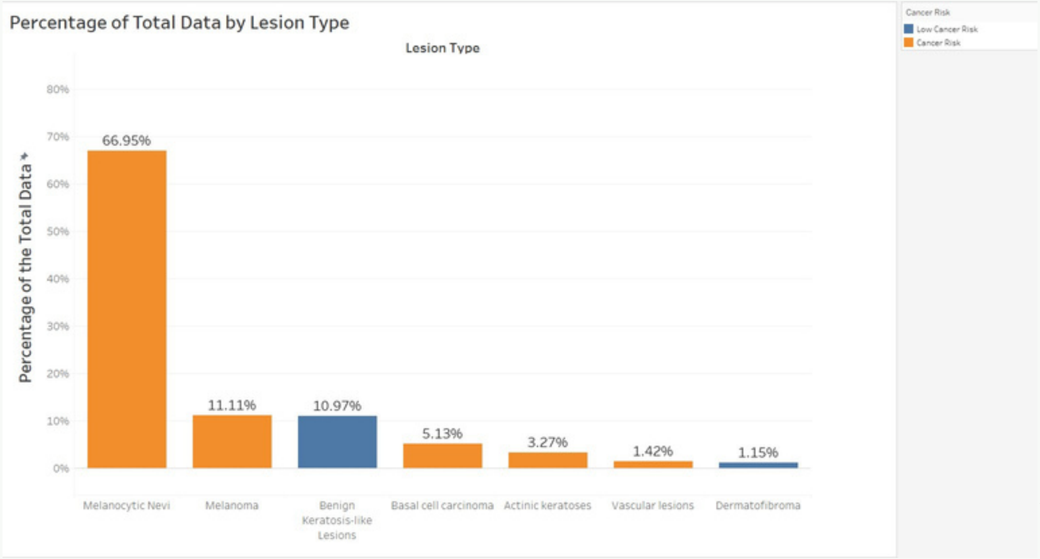

Figure 1

Also worth noting for our data preparation is the conception of our binary variable, "Cancer Risk". Utilizing a number of online medical journals and other publications, this binary variable categorizes lesions as "Low Cancer Risk" if the lesion poses no elevated risk for skin cancer and "Cancer Risk" otherwise. Necessarily, this means that our "Cancer Risk" category is broad, ranging from Melanoma lesions which are the most dangerous form of skin cancer to "Actinic Keratosis", which poses only a marginal risk of developing into cancer. The broadness of this category is by design, as an effective model would do better to air on the side of caution when deciding whether to recommend that a patient seek further care.

Modeling our CNNs

CNNs, or Convolutional Neural Networks, are a kind of neural network frequently used on image datasets. In order to interpret images and to learn to make predictions about the images (in the case of supervised learning), a CNN utilizes a number of layers which begin with default weights and biases, but throughout training, become more and adapted to the training dataset. These layers can come in many shapes and sizes, but CNNs typically have layers for convolution, pooling, dropping, and predicting.

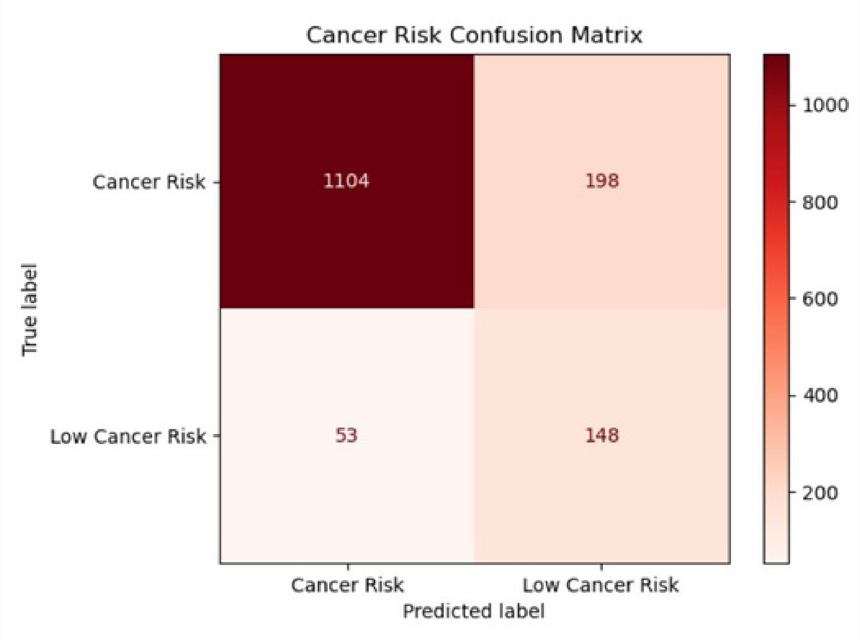

Figure 2

For our binary classification task, our target is whether or not a given skin lesion has an elevated cancer risk with Class 0 having elevated cancer risk and Class 1 being low cancer risk. Since this model dealt with two categories, it utilized a binary classification model (supervised learning). To evaluate our model's performance, we utilize metrics applied on our validation set for each epoch. Throughout modeling, our initial CNN for our binary classification encountered some problems with overfitting. This, we resolved through adding additional dropout layers in our CNN. Transfer learning, or the practice of utilizing layers with weights trained on other image datasets, was considered but eventually decided against, as utilizing entirely trainable weights was found to be the best method for good model metrics in initial testing.

Our final model for our binary classification task, when evaluated based upon its predictions on our testing dataset versus the real testing values, has a precision of ~0.95, a recall of ~0.85, an f1 score of ~0.9, and an overall accuracy of ~83.3%. In order to establish these metrics, "Cancer Risk" has been considered as our model's positive class. As can be seen in Figure 2, a confusion matrix of our final model's predictions on our testing dataset, metrics other than accuracy have been highly influenced by data imbalance in our testing set, which has not been balanced as our training set has. For this reason, the two best metrics for evaluating our model's performance are the overall accuracy and the accuracy at correctly assigning each class. By these two metrics, our model did quite well, but like all models, it has room for improvement.

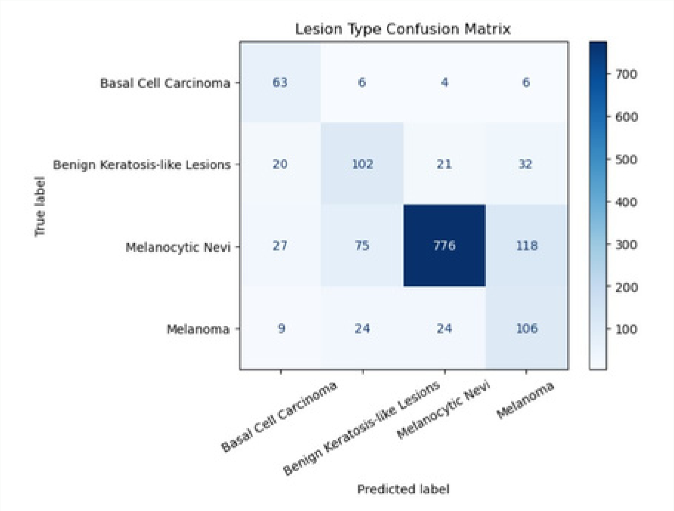

Figure 3

For our categorical classification task, our target is the type of skin lesion is in a given image out of four possible categories. Since this model dealt with four categories, it utilized a categorical classification model (supervised learning). Like with our binary model, model performance for our categorical model is evaluated based upon our validation dataset at each epoch. As transfer learning was not an effective means of training our model for our binary classification, it has not been considered for our categorical classification. Also like our binary model, dropout layers provided an effective tool to prevent overfitting of our training data.

Our final model for our categorical classification task, when evaluated on our testing dataset, had a categorical classification accuracy of ~74.1%. As for our other metrics, Figure 3 features a confusion matrix for our final model applied to our test data. As can be seen this confusion matrix, all categories were correctly predicted the vast majority of the time, but certain categories were more likely to be mistaken for one another than others. One of the most notable examples of this is between "Melanoma" (MA) and "Melanocytic Nevi", (MV) where ~12% of all MV cases were misclassified as melanoma. As with the previous model, this model performs well, but has room for improvement for future iterations.

Findings and Applications

For our binary classification task, the best industry applications for this model is likely to be consumer facing services, such as mobile apps, which allow users to use their technology to either validate or lessen their concerns about skin lesions on themselves and loved ones. These apps could be monetized through crowdfunding, advertisements in the service, or subscription fees.

For our categorical classification task, the best industry applications for this model are likely to be medical professional facing. This model, when utilized by hospitals, medical labs, and other organizations within the healthcare sphere, will allow professionals to seek a second opinion for diagnoses from this model. As has already been noted by the publishers of the HAM10000, models created have already been found to outperform medical professionals.

Next Steps & Future Directions

In order to verify the validity of our models, the most logical next step would be to introduce entirely new data to our models and evaluate our model's performances against that of medical professionals and other existing models. Additionally, throughout this collection of new data, there should be considerations of more categories of skin lesions, including the two categories dropped for our modeling, for future iterations of our model. Additionally, for new data collections, other aspects of the images should be included in our image metadata. For example, future images should have timestamps such that images of the same lesion can be compared over time.

Data Source: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T

Research Sources:
https://www.mayoclinic.org/diseases-conditions/actinic-keratosis/symptoms-causes/syc-20354969
https://www.mayoclinic.org/diseases-conditions/melanoma/symptoms-causes/syc-20374884
https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53