A Convolutional Neural Network Approach for Skin Cancer Detection

Andy Dai, Yi Zhou Tang, Kevin Zhu, and Nicolaus Wong

Western University

May 31, 2020

Abstract

Skin cancer is one of the most common cancers worldwide, and the incidences of skin tumors are slowly increasing globally. While most cases are benign, malignant cases are fatal if the diagnosis is overdue. Expert dermatologists and doctors can accurately identify malignant cancers in their early stages. However, invasive and intensive screening methods are used to diagnose a patient, which increases risks of infections, bleeding, and scarring for the patient. In addition to an increased risk, the hospital bears additional cost and time to complete the screenings.

The aim of this study was to develop an automated diagnosis system that would improve the accuracy of cancer detection. Our method utilized the latest novel advancements in the field of Convolutional Neural Networks (CNN) to detect and assess a set of representative features for skin cancer. Our aim was to create a system that will diagnose patients faster than the traditional methods, be non-invasive, and significantly reduce risks placed on patients. The system provides greater accuracy in skin cancer detection and makes this application an efficient and dependable approach for dermatologists.

Keywords
Melanoma, Skin Cancer Detection, Convolutional Neural Network, Image Classification, Deep Learning

1 Introduction

Skin cancer is one of the most common cancers in North America, with 1 in 5 Americans developing skin cancer by the age of 70. Approximately 12.5 per 100,000 people have melanoma in North America, with cancer statistics predicting 100,350 new melanoma cases by the end of 2020. There are several treatments for melanoma, such as BRAF and PD-1 inhibitors, with a five-year survival rate of 99% if detected early. However, if the melanoma metastasizes, the five-year survival rate would be reduced to 25% [1, 2]. , the number of newly non-melanoma skin cancer (NMSC) patients diagnosed each year is estimated at 5.4 million in North America. Among these 5.4 million cases, over 90% of these patients are diagnosed with Basel Cell Carcinoma (BCC), and the rest with Squamous Cell Carcinoma (SCC) [3, 4]. Although skilled dermatologists are able to detect and remove NMSCs at an early stage through surgery, if SCC metastasizes, they can become lethal since there are only a few effective therapies that have been regulated for advanced cases of SCC. Today, skin cancer is the most common cancer in Canada and the United States [4]. It is vital to have an early skin cancer detection system for all cases, not limited to melanoma, to prevent the development of these cancers to advanced stages, and reduce the number of skin cancer-related deaths in North America.

Using invasive and intensive screenings are a solution for the early detection of skin cancer, but it is not practical for dermatologists or doctors to diagnose every patient using this method. It has been recorded that 83.4% of patients in North America consult their primary healthcare clinics with skin-related disorders before being referred to a dermatologist [5]. In addition, primary healthcare clinics are under pressure to accurately screen patients who present skin-related symptoms and determine which patients are to be referred to dermatologists. Thus, any service or device that can accurately analyze dermoscopic images, and provide the correct classifications of cancer would be beneficial to primary doctors and their patients.

This study aims to determine if training neural networks for automated diagnoses of pigmented skin lesions can aid in increasing diagnosis accuracy for dermatologists.

2 Materials & Methods

In this study, the data primarily used in our analysis and neural net construction was the HAM10000 ("Human Against Machine with 10000 training images") dataset [6] from Harvard Dataverse. The data consists of 10015 dermatoscopic images of common pigmented skin lesions from 2018, collected from different populations. Figure 1 contains a collection of the most common and critical classifications of pigmented skin lesions.

Skin Lesions	Acronyms
Actinic keratoses and intraepithelial carcinoma / Bowen's disease	**akiec**
basal cell carcinoma	**bcc**
benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses)	**bkl**
dermatofibroma	**df**
melanoma	**mel**
melanocytic nevi	**nv**
vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage)	**vasc**

Figure 1: Skins lesions and its acronyms

In addition to the 10,015 images, the dataset also includes categorical features such as age, and gender of each patient. The emphasis of this study is towards the application of image classification in cancer diagnosis, hence only the images were included as features for the models included in this study. However, a detailed summary and analysis of the categorical features are included in the appendix [A] to further understand the data and summarize the characteristics of the demographic.

The data preparations and neural network constructions were completed using the Python packages: *fast.ai*, *pytorch*, *pandas*, and *matplotlib*; the exploratory data analysis was done using Excel.

Figure 2 displays an example image of each cancer classification covered in the dataset.

2.1 Model Architectures

Convolutional Neural Network (CNN) is a type of multi-layer neural network inspired by the

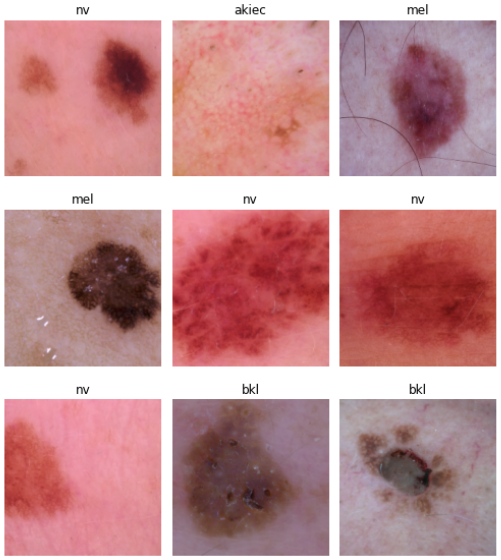

Figure 2: Types of cancers

mechanisms of the optical system of living creatures [7]. In this study, five CNN model architectures are applied and compared for skin cancer detection: AlexNet, VGG-19, ResNet50, ResNet152, DenseNet201.

These models will be trained in two separate stages. The first stage will be trained in ten epochs with parts of its network frozen due to computing resource limitations and to ensure the backward pass does not modify the weights during transfer learning. The second stage will be done by unfreezing the rest of the network with a learning rate specification. As discussed later in this paper, depending on the network and the dataset, the learning rate will exhibit a zero change in its gradient up to a certain point. Then, the loss will gradually increase with an increasing learning rate. By limiting the learning rate to a flat loss rate, the accuracy of the network can greatly increase while simultaneously maintaining a low train and test loss.

2.1.1 AlexNet

AlexNet is the name of a CNN model architecture after Alex Krizhevsky. The network has a similar architecture as the well-known LeNet proposed in [8], the key difference being that it is deeper with more filters per layer with stacked convolutional layers.

In the paper [9], the author presented a new way to parallelize the training of CNNs across multiple GPUs, where the algorithm relies heavily on data parallelism in the convolutional layers and model parallelism in the fully-connected layers.

In 2012, AlexNet outperformed all prior com-

petitors and won the 2012 ImageNet Challenge.

2.1.2 VGG

In 2015, Very Deep Convolutional Network (VGG) was published by [10]. VGG was developed to address the major setback of traditionaL CNNs: depth. The model architecture differs from traditional CNNs in several aspects. The convolutional layers in VGG uses a 3x3 receptive field. The authors proposed that stacking small layers performs better than a single large layer. VGG has three fully-connected layers, the first two layers have 4096 channels while the third layer has 1000 channels, one for each class. Every hidden layer uses the Rectified Linear Unit (ReLU) activation function between the hidden layers.

The proposed approach won the first and second places in the localization and classification tracks of the 2014 ImageNet Challenge.

In our research, the VGG 19-layer model is chosen as one of the models to solve the skin cancer detection problem.

2.1.3 ResNet

In deep neural networks, as the depth increases, accuracy gets saturated and then degrades rapidly along with higher training errors. This is referred to as the degradation problem. Hence, the degradation of the training accuracy implies that not all networks are easy to optimize.

In 2015, Residual Neural Network (ResNet), published by [11], was proposed as a solution to the problem. The network's structure was combined with a residual function to reformulate the layers and provide capabilities for improvements with larger yet less complex networks [12]. Rather than hoping stacked layers will directly fit a desired underlying mapping, the approach explicitly lets these layers fit a residual mapping. The hypothesis was that it is easier to optimize the residual mapping rather than the original, unreferenced mapping.

Let x denote the input of the network. If $H(x)$ is the underlying mapping of data to be fit by a few layers, then the traditional models attempt to use these layers to learn a function $F(x)$ that is equal to $H(x)$. In ResNet, layers learn a residual function, where $F(x) = H(x) - x$, then add x to the results.

The proposed models achieved 3.57% error on the ImageNet test set and the results won 1st place in the ILSVRC 2015 classification task.

In our research, ResNets with 50 and 152 layers were selected to attempt to solve the skin cancer detection problem.

2.1.4 DenseNet

Densely Connected Convolutional Network (DenseNet), is a recent model architecture published in 2016 by [13]. Each layer is connected to every other layer in a feed-forward setting (within each dense block). For each layer, the feature maps of all preceding layers are treated as separate inputs whereas its own feature maps are passed on as inputs to all subsequent layers [13].

DenseNet has multiple attractive characteristics; it allows reuse of features, reduces the number of parameters, and reduces the vanishing-gradient problem.

In our research, DenseNet with 201 layers is selected to attempt to solve the skin cancer detection problem.

3 Results

In stage one, it can be seen in Table 3 that the networks give a wide range of accuracies. These are recorded to compare networks so that the best one can be determined and used in skin lesion classification problems. The best network performance so far is DenseNet201 with an accuracy of 86.4%. As shown later, this network will continue to be the best performing network in stage two. One reason for this is the fact that DenseNet treats inputs separately thereby greatly increasing the network resolution. As a consequence, this network produces the lowest training and testing loss among all the other networks.

Neural Network	Accuracy
ResNet50	0.84623
ResNet152	0.84923
DenseNet201	0.86370
VGG19	0.80330
AlexNet	0.78532

Figure 3: Network accuracy at stage 1, epoch 10

Before stage two, the networks must be assessed in terms of their learning rates. An optimal learning rate range is necessary to ensure optimal training and to reduce training loss. This is achieved by using the learning rate function on the partly trained network. The learning rate is then graphed in order to find the most stable range with little changes to the loss. This is to ensure the network remains stable during training. Figure 4 shows the learning rate graph for AlexNet. It can be seen that the loss remains relatively flat between 1×10^{-5} to 1×10^{-3} and then increases steeply thereafter. This is due to

instability in the network as the learning rate is varied. The learning rate optimizer tests different values for the given network. For any given learning rate, as the name suggests, the network adjusts its training speed and optimizes the network based on the given metric. If it is trained too fast, the resolution becomes over glossed and the loss starts to increase. If the gradient descent is done with a learning rate that is too high or too low, it can easily miss or never reach the local min/max, respectively. Therefore, it is imperative to assess the correct learning rate range before the networks are trained further.

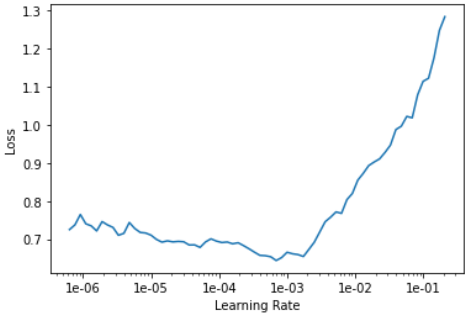

Figure 4: The relationship between AlexNet's learning rate and loss

In stage two, by referring to Figure 5, the accuracies of all five networks improved. In limiting the learning rate range, the models used can reduce their respective losses significantly. In the case of the ResNet neural networks, their accuracies saw very little improvement. This is because, for this particular network type, although it exhibited a stable learning rate loss over a certain range of rates, that range was not wide enough to see significant changes in the model training accuracy. Therefore by limiting the learning rate, the limited change of the gradient of the learning rate/loss function will result in an accuracy that seldom changes, even with more epochs. The learning rate plots of ResNet50 and ResNet152 are shown in Figure 6 and Figure 7.

Neural Network	Accuracy
ResNet50	0.85921
ResNet152	0.85921
DenseNet201	0.89416
VGG19	0.85572
AlexNet	0.83675

Figure 5: Network accuracy at stage 2, epoch 5

From stages one and two, the best network has been identified as DenseNet201 with an accuracy of 89.4%. Intuitively it makes sense as

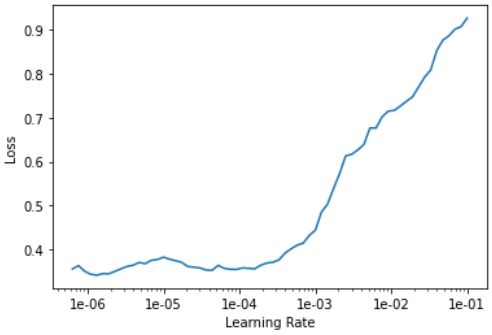

Figure 6: The relationship between ResNet50's learning rate and loss

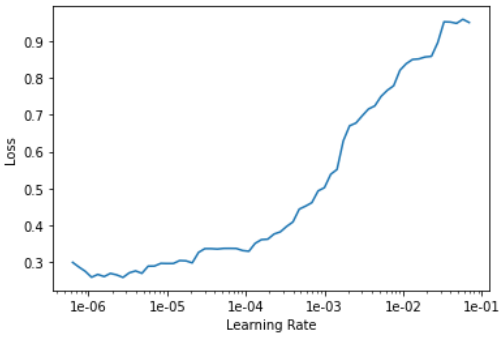

Figure 7: The relationship between ResNet152's learning rate and loss

this particular DenseNet network is much larger than the other networks. With 201 hidden layers, the network is capable of identifying more inputs which will result in improvements in the network resolution. A higher resolution combined with a partly fitted model and an optimal learning rate resulted in a 3% increase in the accuracy of the network.

A graphical representation of the results achieved are shown in Figure 8 and Figure 9. It can be seen from both neural networks that the best predictions occur for *melanocytic nevi (nv)*. This is because this skin lesion type represented the majority of the dataset and will result in better accuracy for that specific label as compared to the rest of the labels.

Another method of analyzing the results is by looking at the top loss plots in Figure 10. Here, the images are labeled in the order: predicted label, actual label, loss, and the probability of actual class. The reason neural networks make mistakes with these specific combinations is because of similarity. As accurate as DenseNet201 can get, it is not 100% accurate as information gets lost during backpropagation. As such, the lesions that have the potential to look the most similar: *melanoma (mel)*, *melanocytic nevi (nv)*, and *benign keratosis-like lesions (bkl)* are mis-

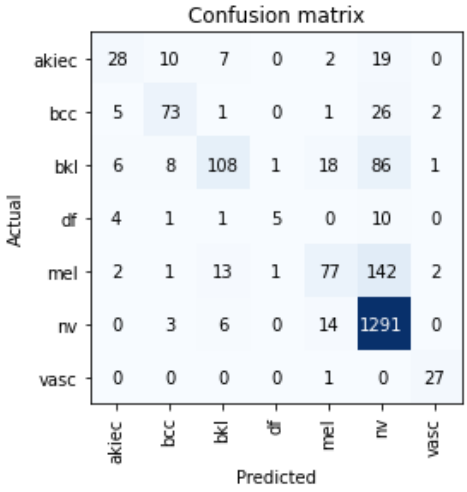

Figure 8: AlexNet

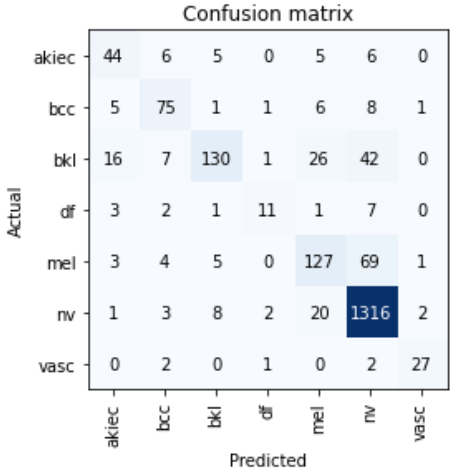

Figure 9: DenseNet201

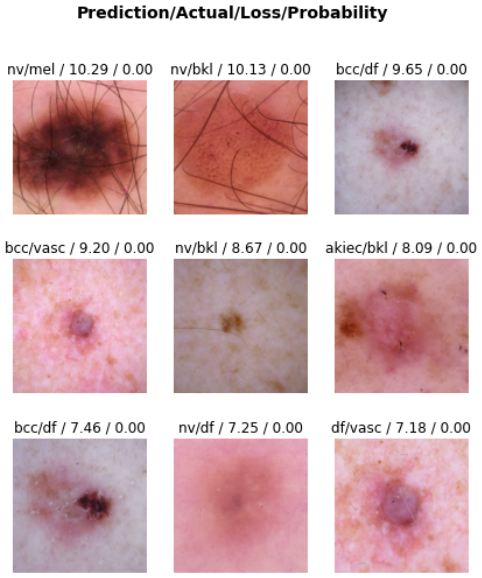

Figure 10: Plot of top 9 losses of DenseNet201

Actual	Predicted	Occurrences
mel	nv	80
bkl	nv	46
nv	mel	25
bcc	nv	17
bkl	mel	17
mel	bkl	15
akiec	nv	12
nv	bkl	12
akiec	bkl	11

Figure 11: DenseNet201 top misclassifications by skin lesion type

classified by the network.

Figure 10 is also presented in Figure 11 for conciseness. The number of occurrences is also shown as this information could be used to flag potential misclassifications if this model is actually deployed.

4 Discussion

By labeling the data by skin lesion type, using five different convolutional neural networks for modeling and assessing the worst performances of models, we were able to achieve a model accuracy of 89.4% with DenseNet201. This model can be confidently used in medical services to classify skin lesions. Realistically, there are thousands of skin lesions, but due to limitations in the dataset and computing resources, we were only able to use this dataset to train the model with a batch size between 16 and 32. The batch size is the number of images the network sees at a certain given time. The higher the batch size, the higher resolution the network can achieve. Ideally, it would be trained using a batch size of 64 or 128 with thousands of labels so that this model can be deployed for a wide range of lesion classifications.

Lastly, there is potential for overfitting as the training loss drops below the validation loss at the 7th epoch. To combat this problem, more data is needed for the network to train on so that a better validation loss can be achieved. It is impossible to eliminate the problem of overfitting, but it could be mitigated to a point where it is bare negligible.

Applications of artificial intelligence in the field of cancer diagnosis has been a popular topic among many researchers. In [14], a skin lesion detection system was developed using Google's pre-trained CNN model Inception-v3, which was able to achieve an overall accuracy of 65.8% for the validation set. In [15], the proposed CNN

architecture for skin cancer detection showed an accuracy of 80% in training and 78% in testing; the process used a total of 50 epochs and a learning rate of 1×10^{-3}. For the three-part competition organized by the International Skin Imaging Collaboration (ISIC) in 2017, a transfer learning method was proposed by [16] for *melanoma* detection. The model was able to produce an 83% test accuracy for distinguishing *melanoma* from nevus and *seborrheic keratosis* and a 92% for distinguishing *seborrheic keratosis* from *nevus* and *melanoma*.

Conclusion

By using convolutional neural networks as a method of image classification, we have been able to train a model using DenseNet201 on the HAM10000 dataset with an accuracy of 89.4%. Due to the way DenseNet passes inputs forward into subsequent layers, the resolution achieved via this network is far more superior than the other networks. By separating the training into two stages, the learning rate optimizer was able to use the partly trained network to find an optimal range for the learning rate. As a result of this crucial step, the accuracy achieved after the fifth epoch in stage two produced results that were far superior than the previous stage and previous attempts in classifying this dataset.

As such, this model can be deployed into beta to assess the viability of its use. In hospitals and clinics, skin lesions like *melanoma* and *melanocytic nevi* can be classified and diagnosis with a much higher accuracy which can ultimately increase the efficacy of such clinics. It should be noted that due to limitations in computing resources, time, and data, we were not able to train the network past 89.4%. However, if given more resources and time, a 90%+ accuracy could potentially be achieved.

Computing weights for any deep learning network requires large amounts of computational power. The models were trained locally and the training batch size was limited to prevent overheating and reduced computation times. However, after a certain number of epochs, the GPU still ran out of memory. To compensate for this, the script and data were moved and trained with a cloud GPU.

At this time, the batch size was still limited. As such, the accuracy is expected to be much higher for all the trained networks if the batch size was increased to 64 or 128. Lastly, the dataset obtained had only 10015 rows of data with four distinct features that were deemed significant in the exploratory data analysis. If a larger dataset with more labels and features were used, the training time would increase, but the accuracy of the model would also increase. A potential goal for improvement would be to use a higher batch size with a bigger dataset so that the accuracy of the network can improve, ideally beyond 90%.

Methods of image classification, semantic segmentation, Kalman filters, and Bayesian inference have enabled modern advancements in fields of medicine, self-driving cars, finance, and so much more. Convolutional neural networks like ResNet and DenseNet are only the beginning of what could become the biggest field of research in the near future.

Acknowledgements

We would like to express deep gratitude to Anish Verma at STEM Fellowship for valuable suggestions in the development of this work.

References

[1] Melanoma: Statistics. urlhttps://www.cancer.net/cancer-types/melanoma/statistics/, 2020.

[2] Melanoma: Types of treatment. urlhttps://www.cancer.net/cancer-types/melanoma/types-treatment, 2019.

[3] The American Cancer Society medical and editorial content team. Key statistics for basal and squamous cell skin cancers. urlhttps://www.cancer.org/cancer/basal-and-squamous-cell-skin-cancer/about/key-statistics.html/, 2020.

[4] Skin cancer facts statistics. urlhttps://www.skincancer.org/skin-cancer-information/skin-cancer-facts/, 2020.

[5] Chris van Weel Peter C M van de Kerkhof Piet Duller Pieter G M van der Valk Henk J M van den Hoogen J Hans J Bor Henk J Schers Andrea W M Evers Elisabeth W M Verhoeven, Floor W Kraaimaat. Skin diseases in family medicine: Prevalence and health care use. urlhttps://pubmed.ncbi.nlm.nih.gov/18626035/, 2008.

[6] Philipp Tschandl. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018.

[7] Farhana Sultana, Abu Sufian, and Paramartha Dutta. Advancements in image classification using convolutional neural network. *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, Nov 2018.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In *Intelligent Signal Processing*, pages 306–351. IEEE Press, 2001.

[9] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks, 2014.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2015.

[12] Yaser Saleh and Ghassan Issa. Arabic sign language recognition through deep neural networks fine-tuning. *International Journal of Online and Biomedical Engineering (iJOE)*, 16:71, 05 2020.

[13] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016.

[14] P. Mirunalini, Aravindan Chandrabose, Vignesh Gokul, and S. M. Jaisakthi. Deep learning for skin lesion classification, 2017.

[15] Ardan Adi Nugroho, Isnandar Slamet, and Sugiyanto. Skins cancer identification system of haml0000 skin cancer dataset using convolutional neural network. *AIP Conference Proceedings*, 2202(1):020039, 2019.

[16] Dennis H. Murphree and Che Ngufor. Transfer learning for melanoma detection: Participation in isic 2017 skin lesion classification challenge, 2017.

A Appendix

A.1 Exploratory Data Analysis of Categorical Features

In the analysis of the HAM10000 dataset, there are four additional categorical features provided in addition to the images of pigmented skin lesions. The features are sex of the patient (*sex*), age of the patient (*age*), localization of the image (*localization*), and the examination type (*dx-type*) done for each patient.

In addition to constructing a neural network, it is crucial in the world of data and statistics to understand the underlying characteristics of the population represented in the data.

In an exploratory analysis of the data, we observed that more cases of skin cancer tend to be male, at 54.29%, compared to females making up 45.71%. Figure 12 displays the population divided via the genders from the data. A look at the ages of the patients shows us that the mean age of patients is 51.53051, with a skew towards the upper ages. Figure 13 displays the population distributed among the ages. Taking a closer observation at the distribution of ages, we can see that there is a small but significant increase of diagnoses around the age of 5. We can see that there is a greater concentration of patients around the mean, with a 95% confidence interval of *(51.53, 52.20)*, indicating that 95% of the time, the age of a patient will be between 51.53 and 52.20.

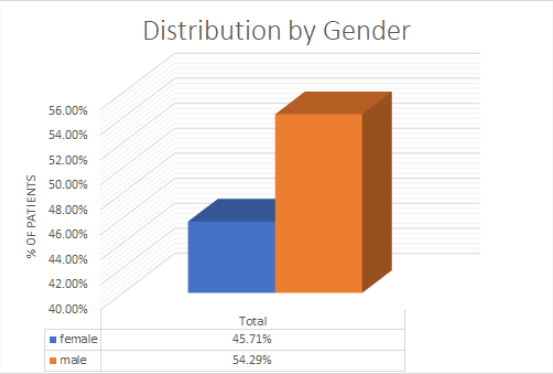

Figure 12: Distribution of patient genders

Investigating the diagnosis aspect of the data, we can see from Figure 14 that the most common diagnosis made by a doctor was for *melanocytic nevi*, *nv*. Since *melanocytic nevi* has no known methods of prevention, it is crucial to screen potential cases often to ensure early treatment.

The localization of cancer is very important. It is crucial to understand which areas more frequently get cancer, to be able to recommend more preventative measures. A look at our data

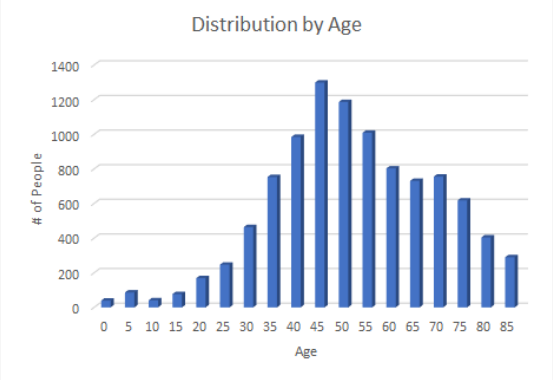

Figure 13: Distribution of patients by age

distributed by localization tells us that most commonly, skin cancer attacks the back and lower extremities, making the posterior side the most active for skin cancer. This may be due to individuals facing away from the sun since UV ray exposure is one of the main factors of skin cancer. Our dataset did not contain additional information regarding environmental factors, but it would be interesting to extend the research to include such factors.

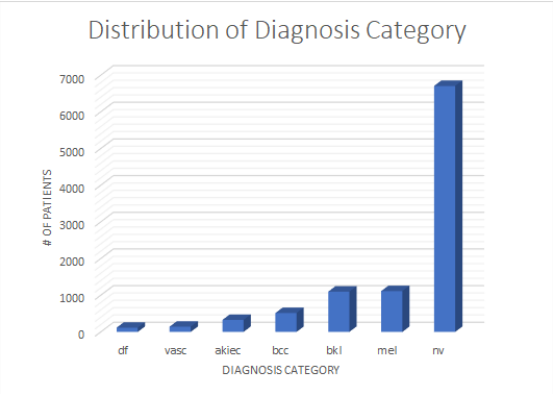

Figure 14: Distribution of by the type of cancer

Similar to understanding the localization of cancer, it is as important to investigate how cancer is diagnosed. Looking at Figure‘16, the data tells us that a significant majority of diagnoses stem from histopathology. Since histopathology requires tissue samples to be examined under a microscope, the tissue is usually removed by cutting it from the patient leading to higher risks of infections and other illnesses.

In statistics, it is important to understand the effect of interactions between features and how multiple features combined can bring more insight than a single feature. Diving deeper into the data, we can see that when we arrange the data by *age* and *sex*, we can see from Figure‘17 that females tend to have skin cancer more often prior to their fifties, with male cases of skin can-

cer significantly greater than females after the age of 50.

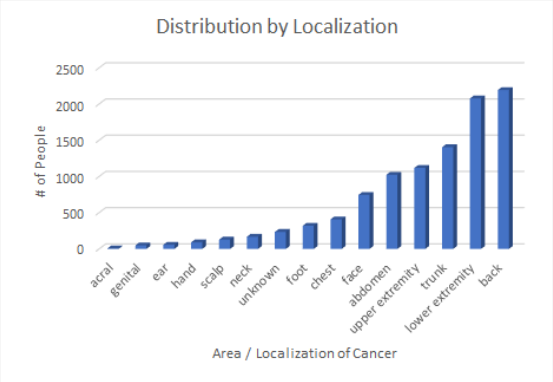

Figure 15: Distribution of patients by the localization of the diagnosed cancer

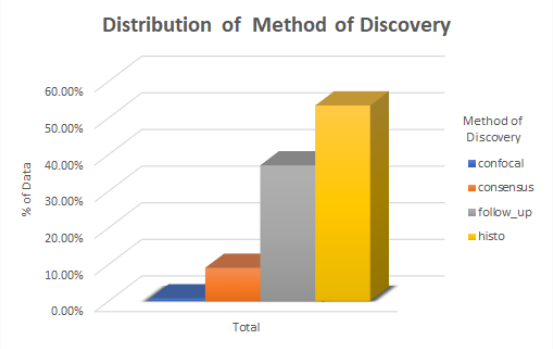

Figure 16: Distribution of the diagnosis method among patients

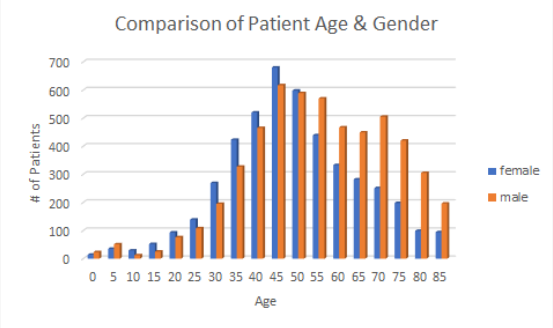

Figure 17: Distribution of patients by age

When we arrange the data by *localization* and *sex*, we can see from Figure 18 that females are significantly more common in getting acral cancer, and skin cancer in the hands, and genitals. Males are more common in getting skin cancer on the scalp, chest, and back.

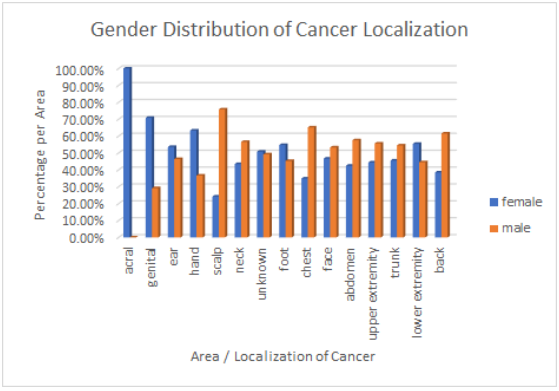

Figure 18: Comparison of the localization of cancer between genders

A.2 Feature Ranking

To understand the implications of the categorical features, two feature importance techniques were used to evaluate and rank the features: random forest feature importance and permutation importance.

A.2.1 Random Forest Feature Importance

Random forest is a non-linear statistical learning method useful for modeling complex relationships between features and target variables. It is an ensemble of tree models such that each tree depends on the values of a random subset of the original data with the same distribution for all trees in the forest. One attribute of the model is that it has an impurity-based feature importance method - Mean Decrease in Impurity (MDI).

In this study, all four categorical features were fitted with a random forest model with 250 trees to predict the target variable. The feature importance is derived from the random forest model for ranking purposes. Figure 19 illustrates the MDI values of each feature. It can be seen that, in order of importance, all three variables age, examination type, localization showed significant results with age being the most volatile, hence the most significant.

A.2.2 Permutation Importance

To further validate the rankings based on MDI, permutation importance is applied to the same random forest model for each feature. To compute the permutation importance of a feature, the accuracy of the original model is calculated and recorded first, then the particular feature is permuted and the same metric is calculated and recorded again. The difference between the met-

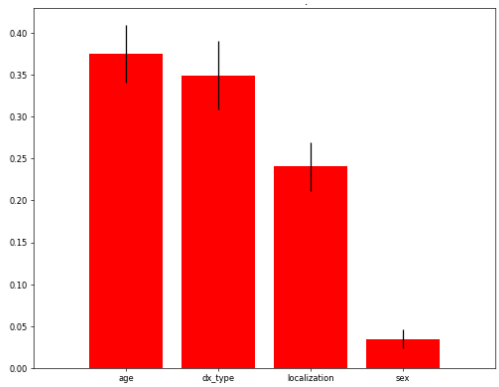

Figure 19: Variable importance based on Random Forest.

ric before and after permuting is defined to be the permutation importance.

Figure 20 shows the permutation importance of each feature. It can be observed that all three variables age, examination type, localization showed significant results. Examination type achieved the highest importance, while sex remains to be the least significant feature among the four.

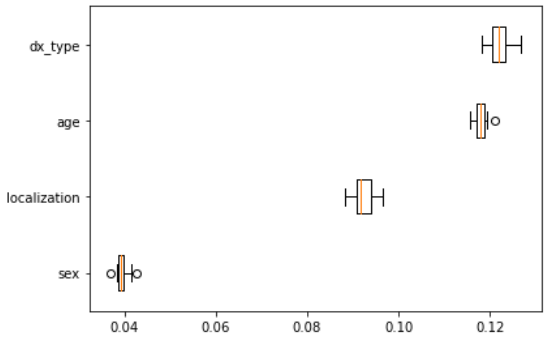

Figure 20: Permutation importance of each variable.