Diagnosing Skin Cancer using Machine Learning Algorithms and Easily Available Information

Oliver T. Midbrink,
Viktor Rydbergs Gymnasium 2020

Abstract

In this work, machine learning models developed using easily available information for diagnostics of skin lesions will be investigated. The goal is to see if making diagnoses is possible using the ISIC Archive and a VGG16 neural network. The model is fed a dermoscopic image of a skin lesion, then categorizes the image into one of the 18 diagnoses available. As the results show, this method of diagnosis significantly outperforms the act of randomly selecting a category. The best attempt, out of five in total, outperforms random diagnostic accuracy by a factor of 5.4, reaching an average of 44.5% of predictions correct on the first try. Although this method does not perform as well as educated dermatologists or state-of-the-art neural networks, it gives an indication towards the right diagnosis. Thereby, serving the purpose of this study by showing that a machine learning model capable of diagnosing skin lesions can be created using information from sources such as the internet. An application of this software could be, for example, an app that the average person can use when investigating a suspicious skin lesion on their body.

Contents

1. Introduction

In today’s society, many people have heard the term AI but fewer understand what it actually means. AI stands for artificial intelligence and most of the AI software used today falls into the subcategory of machine learning, which was first named by Arthur Samuel in 1959 at IBM. Today, AI is being used in everything from optimizing marketing campaigns to enabling self-driving cars and one very important use case is medicine. Here, machine learning is used to detect lung cancer, structural abnormalities and save lives. Since skin cancer can be extremely dangerous, and the rates of skin cancer is rising in the world, it is important to explore the possibility of using AI in this field. The study will try to see if it is possible to diagnose skin lesions using a machine learning algorithm, and in the case that it is possible, get an indication of what performance is to be expected. An attempt will be made to answer the following question: How accurately can a machine learning algorithm developed using free, easily available information diagnose skin lesions?

2. Background

2.1 Skin Lesion

A skin lesion is an abnormality in color or texture in a patch of skin, that differs significantly in appearance from the surrounding area of skin. Examples of skin lesions are birthmarks, a rash, or in some cases melanoma. A type of skin cancer. There are many different types of skin lesions and this study uses a dataset containing 18 different types. These can be seen in the results-section on page 10.

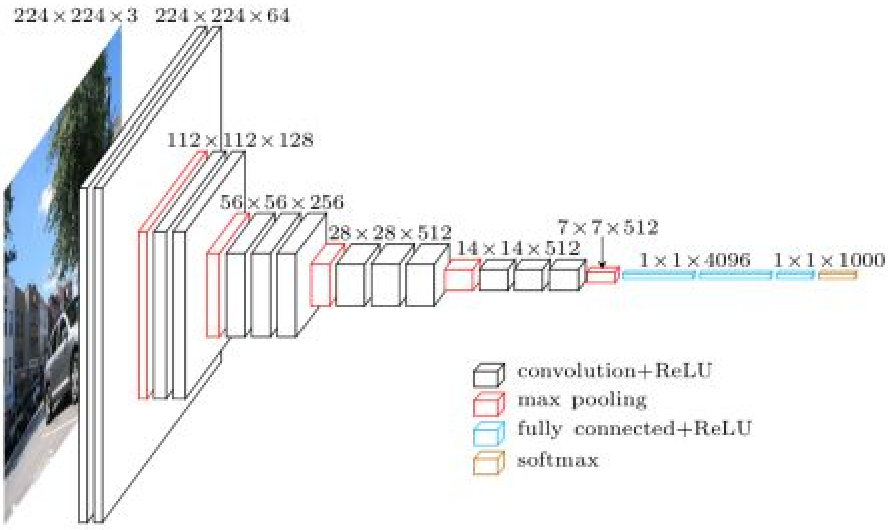
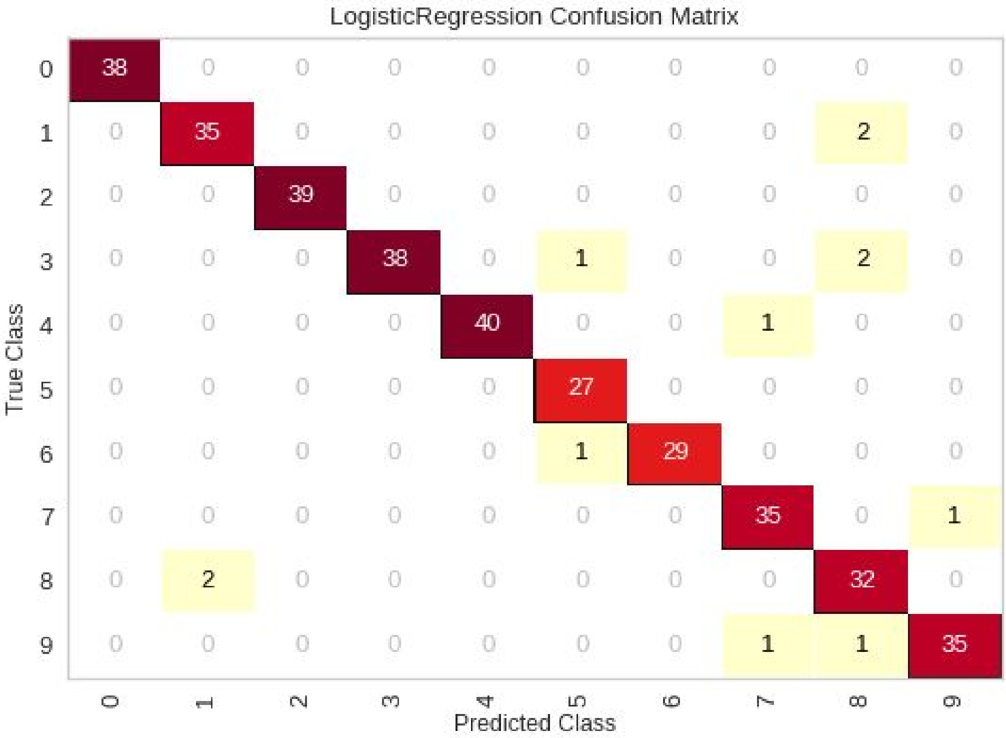


Figure 1. Diagram of the VGG16 architecture.

2.2 VGG16 Explained Briefly

This architecture was invented in 2014 by a group at Oxford University. Although not as advanced as state-of-the-art architectures, it set the foundation for image classification architectures. The first part of the VGG16 architecture consists of convolutional layers. These layers are very well suited for image classification since they are good at recognizing features such as edges and lines. As the number of convolutional layers increases, the neural network is able to recognize more advanced shapes. The second part consists of fully connected layers that are used to interpret what these features mean for the prediction and also which features are important or unimportant for the task. Lastly, there is a type of fully connected layer called the softmax layer. This part makes the final decision of what the network outputs. Overall the performance of this architecture was very high for its time and still is quite good today. Since it is relatively simple compared to modern architectures, it is used in this project.[1]

[1] Karen Simonyan and Andrew Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2015, https://arxiv.org/pdf/1409.1556.pdf



	0	1	2	3	4	5	6	7	8	9
0	38	0	0	0	0	0	0	0	0	0
1	0	35	0	0	0	0	0	0	2	0
2	0	0	39	0	0	0	0	0	0	0
3	0	0	0	38	0	1	0	0	2	0
4	0	0	0	0	40	0	0	1	0	0
5	0	0	0	0	0	27	0	0	0	0
6	0	0	0	0	0	1	29	0	0	0
7	0	0	0	0	0	0	0	35	0	1
8	0	2	0	0	0	0	0	0	32	0
9	0	0	0	0	0	0	0	1	1	35

Figure 2. Confusion matrix showing accurate performance.

2.3 Understanding Performance Using a Confusion Matrix

A confusion matrix is a square shaped grid or matrix that is used to get an overview of the performance of a prediction algorithm, in this case a machine learning model. The columns and labels on the bottom of the matrix are categories that the model predicted. Rows are the true values. The values inside each cell in the grid represents how many times the model predicted the category of the column, when in fact the true category is that of the row. The model is only correct when the cells have the same label on the left and at the bottom. As a result, the confusion matrix should form a diagonal line from the top left to the bottom right, as seen in figure 2, if it is very accurate. If the model predicts the same for two categories, it could point to similarities between them. In contrast, different outputs for the same input could mean the inputted category has high variation. These facts are useful to know when viewing the results of this study.

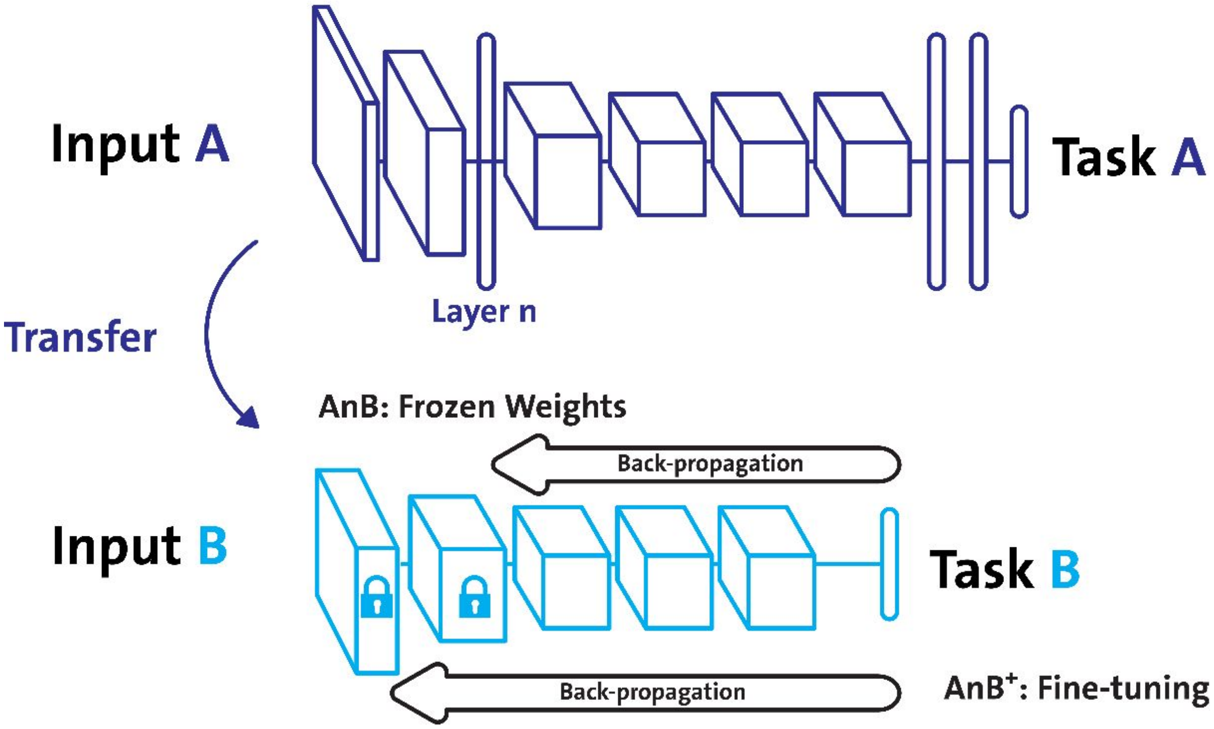


Figure 3. Illustration of the principle of transfer learning

2.4 Transfer Learning

A common problem in the field of machine learning is a lack of data. The result is that the model simply does not have enough information in order to be able to make a generalisation of the problem. Also, training neural networks can take a lot of time. To solve these issues, the practice of transfer learning was invented. A simple explanation of how transfer learning works could be: the practice of using experience (trained parameters) from one task to make better predictions in another task. In the case of skin lesion classification there is quite little data to use, however some of the aspects of skin lesion classification are quite common. For instance recognizing edges, textures and shapes. Therefore, it is possible to use the parameters obtained through training on a large dataset like imagenet, and transfer these to the model. The model is then fine tuned by training on the ISIC dataset. This gives the model a foundation to start from, since it does not have to start from a random configuration, and it can also result in the model being able to see patterns it might not otherwise have detected. In this study, the model uses weights from the imagenet dataset and freezes the first 8-14 layers in order to not interfere with these parameters.

3. Method

The method used for diagnosing skin lesions consists of several parts. Major parts that are used in the end product are two convolutional neural networks that have separate functions. One is used to crop a high resolution image in order to prepare high quality data, the second is used to diagnose the skin lesion. The end product is a function where a path to a high-resolution image is inputted and a diagnosis is outputted. However, due to the scope of this study, the locator network will not be discussed in detail. It is mostly used to preprocess and standardize the data and has less to do with the actual classification.

3.1 Method Summary

During the study, a number of different trials will be made in order to approach the problem from different perspectives. Methods and techniques mentioned below are used in some trails, although not necessarily in every trial. Depending on the trial, either class weighting or the sequence generator is used. Both are not active at the same time. The goal of each trial is to achieve as high categorical accuracy as possible on the validation data. Important to note is that the model will only be exposed to the test data when the trial is finished and does no longer produce any improvements in terms of higher validation accuracy. This way the test data does not directly or indirectly contribute to the improvement of a trail, thereby remaining able to objectively determine the performance. Even though the model does not directly use the validation data for training, the developer does use the information gathered from tests on the validation data for hyperparameter adjustments, which in turn increases performance of a model. Therefore the validation data is biased and will not be used to determine results in this study.

3.2 Data

The data used for this project is provided the ISIC archive. This dataset contains around 25000 high-resolution images with metadata for each image and segmentations for around 13000 of these images. In the metadata, information such as clinical diagnosis, position on the body, age, gender and more are included. However, only the clinical diagnoses, images, and segmentations are used in this study. [2]

3.3 Hardware and Software Used

The development of the machine learning software in this study has leveraged the community support and flexibility of the keras machine learning framework. Python 3.6 was the programming language used, running in a custom Anaconda environment on a windows 10 system. The hardware used consist of a 6th generation i5 processor for data augmentation and training of models was done on a Nvidia CUDA compatible GTX 1060 graphics card.

3.4 Localization

Part of the software used is a localization neural network of VGG16 architecture. The function is to localize the skin lesion and crop a small image around it as seen in figure 4. During the preprocessing of the ISIC dataset, all images are fed through this software in order to standardize the size of images. Also, since the resolution of the classifier is only 224 by 224 pixels, this helps to obtain as much relevant detail as possible.

[2] International Skin Imaging Collaboration, *ISIC Archive,* 2019, https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main

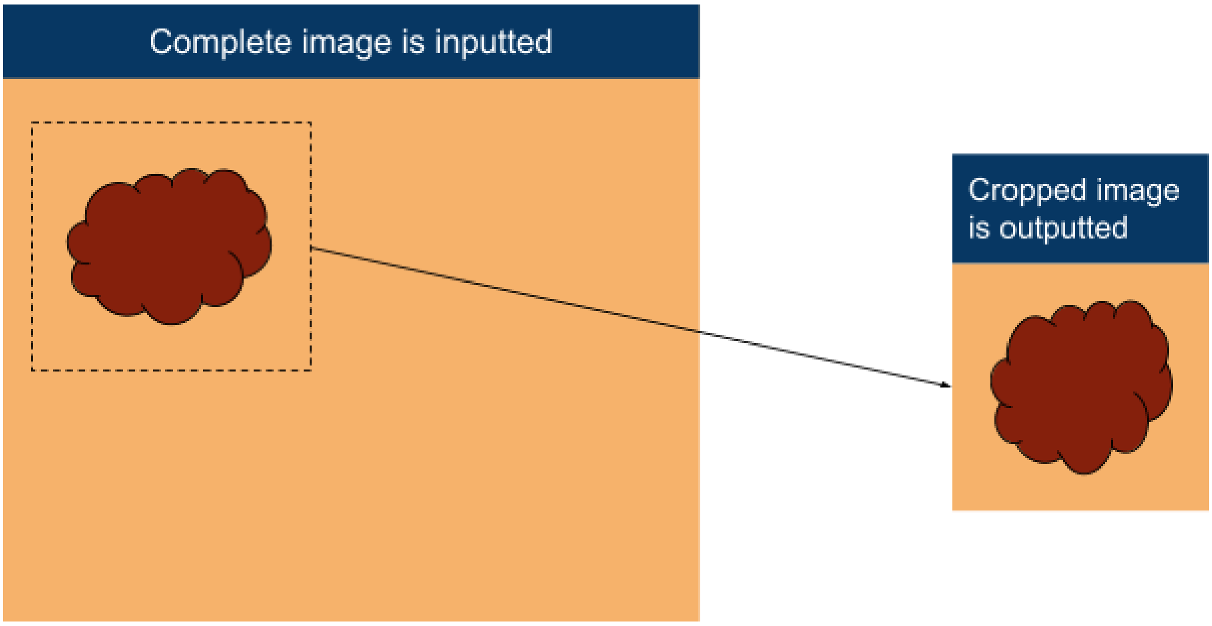


Figure 4. Diagram showing the function of the localizer network. A large image is inputted and a 224x224 pixel image is outputted. In the output image, skin lesions are placed in the center with a 10 pixel padding on each side.

3.5 Classifier

The second part contains a standard VGG16 classifier (see figure 1) that has been pretrained on imagenet, a large image dataset. To adapt the architecture for this particular use case, some of the layers have been retrained on the ISIC data. Also, the fully connected layers have been modified to fit the data format. Mainly the number of outputs.

3.6 Data sequence generator

During the training phase of some trails, a custom data sequence generator has been used. The categories (diagnoses) have highly varying number of images in the training data. Therefore, a method is used to ensure batches of training data contain an even distribution of categories. This is to resist the classifier from making biased predictions toward the larger categories. The first images from category 1 through 18 are fed sequentially, giving a sequence of images with increasing category number. Then, the second image from category 1 through 18 etc. This is illustrated in figure 5 below. If the algorithm has iterated all images from a category, the category is repeated, as seen in category 2, 5 and 7 below. If a category contains no data, this category is skipped as shown in column 6 in figure 5. The arrow at the bottom of figure 5 shows that when all data has been fed, a new epoch is started and the sequence generator loops over the data once more.

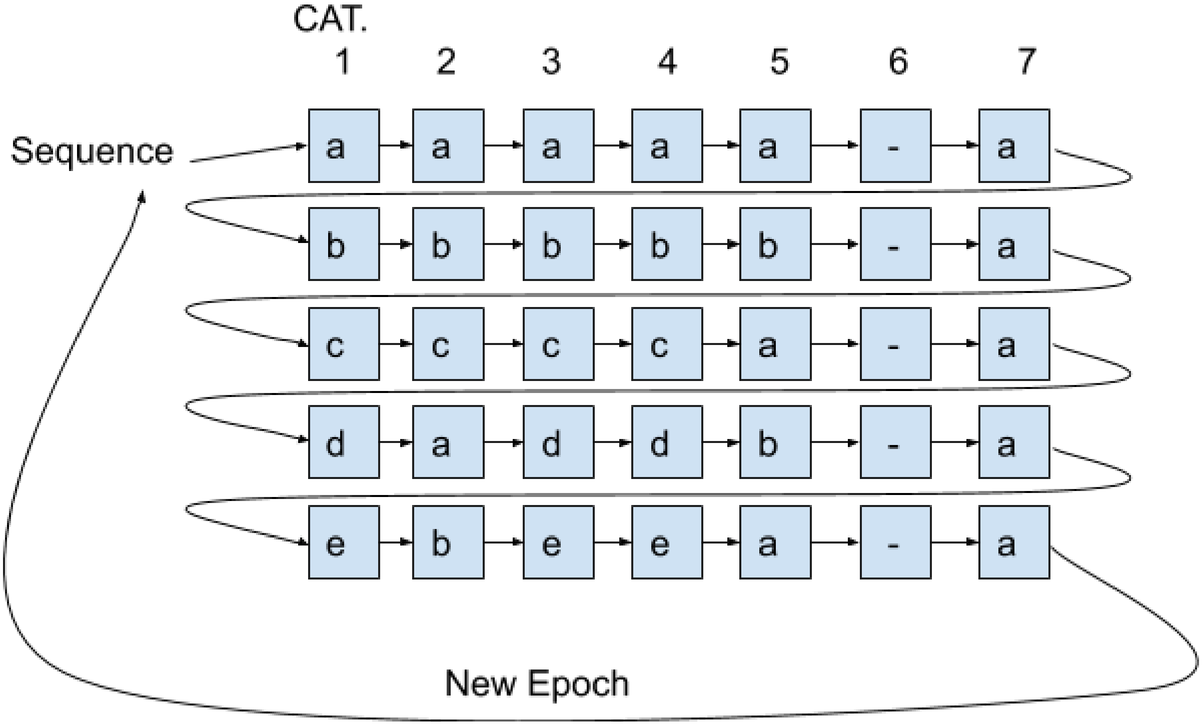


Figure 5. Showing in what order data is fed to the model during training.

3.7 Data Augmentation

Since the dataset is relatively small, data augmentation is also used. Whenever an image is fed to the classifier model, there is a chance the image will be flipped vertically or horizontally. Additionally, the image is translated up to 50% of the width or height in any direction and the brightness and contrast is changed randomly. This prevents overfitting and makes the generalization process easier for the model.

3.8 Class weighting

This is an alternative method to the sequence generator that is useful when dealing with an unbalanced frequency of classes or categories (diagnoses). In this case, one of the 18 classes represents 78% of the total number of training images. If not using class weights or the sequence generator, the model would be biased toward this class since it would be fed mostly this type of data during training. The model would blindly output the same value no matter what type of input is showed. This problem is combatted by class weigths and the sequence generator, which forces the model to learn correctly. Class weighting means that during the training process, some errors are valued as more important than others by the loss function. In this case, the underrepresented classes have higher class weights which makes a prediction error when inputted this type of data much higher. On the other hand, if the model predicts incorrectly when fed the common type of data it is much less important. When using balanced class weights, this can effectively combat uneven data. Additionally it is possible to manually adjust these weights for a particular purpose or use case in order to change the behaviour of the model. In this study, the weights have been adjusted to some extent in order to be biased towards dangerous types of skin lesions. It is more safe to overrepresent the danger of a inputted skin lesion, than to underrepresent it. If the risk level is underrepresented, it could potentially lead to treatment of the skin lesion being deferred or avoided. To reduce this error, the class weights of dangerous types of skin lesions were increased by 70%. Two classes contained only one datapoint and corresponding class weights were reduced dramatically from 834 to 30 to prevent overfitting. The non-dangerous or benign types of skin lesions had their weights reduced by 40%.

3.9 Calculating Accuracy, Specificity and Sensitivity

When evaluating the accuracy of the model (trained neural network), the model is shown all images of each category sequentially. For each category the classifier model makes predictions for all images of the category and the percentage of correct predictions is calculated. When it has iterated through all categories, the code performs an unweighted average for the percentage of correct predictions from each categories. This method can be used to get an overview of the classifier model's performance. Furthermore, specificity and sensitivity are also important when evaluating performance. Calculations for these are listed below:

$$Sensitivity = \frac{True\ Positive}{Positive} \qquad Specificity = \frac{True\ Negative}{Negative}$$

The term true positive refers to the number correct prediction where the attribute, for instance melanoma, is present. True negative signify correct predictions for when the attribute is not present. Positive and Negative refers to the total number of positive and negative data elements.

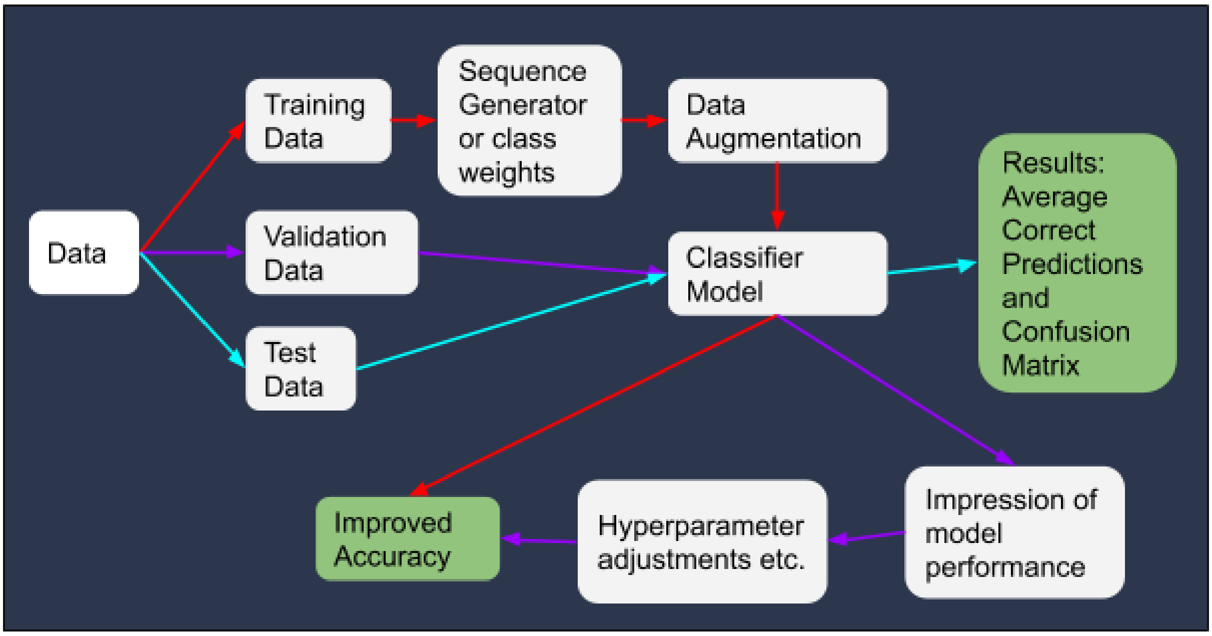

Training Data
Sequence Generator or class weights
Data Augmentation
Data
Validation Data
Classifier Model
Results: Average Correct Predictions and Confusion Matrix
Test Data
Improved Accuracy
Hyperparameter adjustments etc.
Impression of model performance

Figure 6. Diagram of data flow when training and testing the classifier model.

3.10 Data Flow

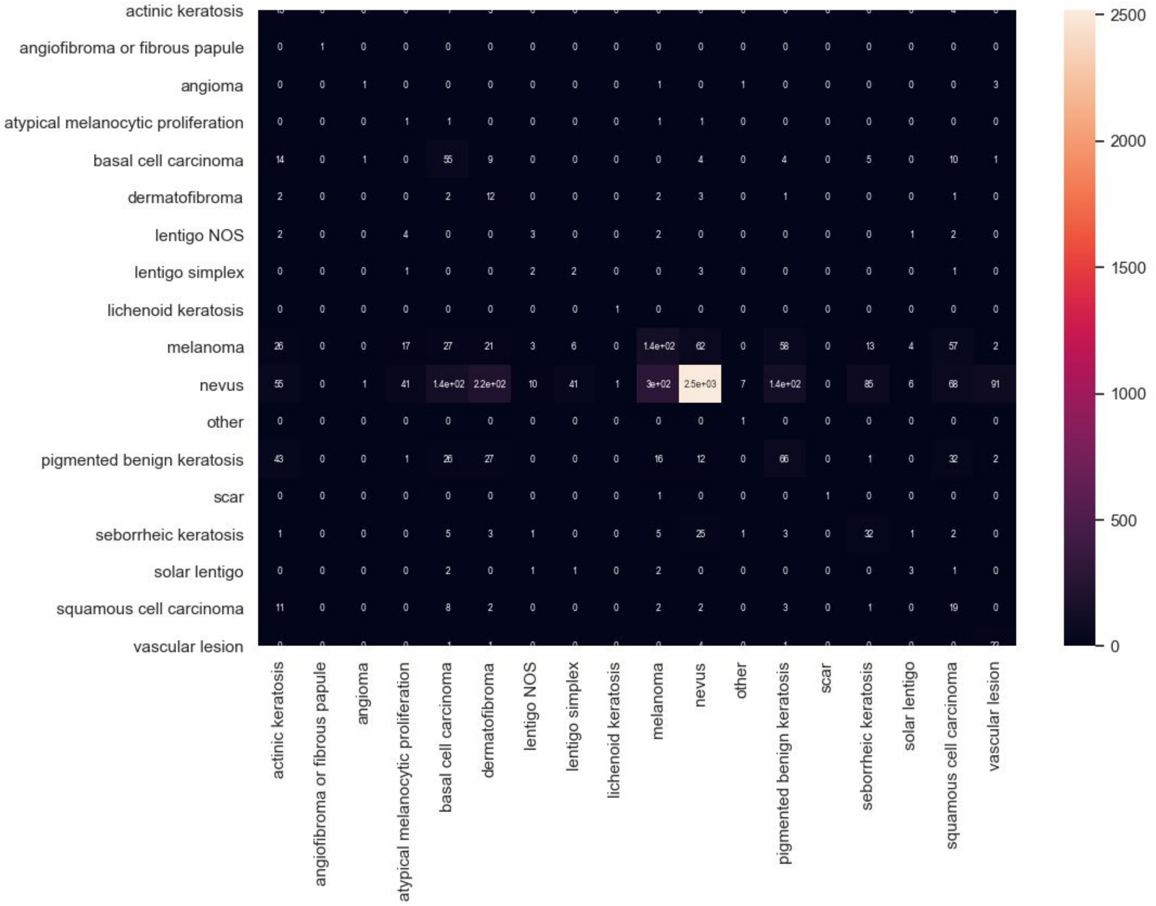
In order to produce the diagnosing function described at the top of the method section, both the localizer model and classifier model was trained for several hours using data derived from the ISIC Archive. In figure 6 above, an overview of how the classifier model was trained and tested can be seen.

4. Results

In order to refer more easily to the different trails, they have been given names describing configuration and a trail number. Seq-X means that the sequence generator is enabled and that class weights are not used. The X identifies in which order the trails of this configuration has been performed, for example Seq-1 refers to the first trial with the sequence generator. Wei-X means that the sequence generator is disabled and that adjusted class weights are used instead. Also, to clarify what accuracy refers to in this context, as opposed to the context of binary classification, the term refers to how many percent of the test images of a certain that were correctly predicted on the first try, similar to sensitivity or categorical accuracy. See part 3.9 for a more detailed explanation.

Diagnosis	Diagnosis id	Frequency
actinic keratosis	0	22
angiofibroma or fibrous papule	1	0
angioma	2	5
atypical melanocytic proliferation	3	3
basal cell carcinoma	4	102
dermatofibroma	5	22
lentigo NOS	6	13
lentigo simplex	7	8
lichenoid keratosis	8	0
melanoma	9	431
nevus	10	3737
other	11	0
pigmented benign keratosis	12	225
scar	13	1
seborrheic keratosis	14	78
solar lentigo	15	9
squamous cell carcinoma	16	47
vascular lesion	17	28

Figure 7. Table showing a description and number of test images for each diagnosis.



	actinic keratosis	angiofibroma or fibrous papule	angioma	atypical melanocytic proliferation	basal cell carcinoma	dermatofibroma	lentigo NOS	lentigo simplex	lichenoid keratosis	melanoma	nevus	other	pigmented benign keratosis	scar	seborrheic keratosis	solar lentigo	squamous cell carcinoma	vascular lesion
actinic keratosis	[illegible]	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	4	0
angiofibroma or fibrous papule	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
angioma	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	3
atypical melanocytic proliferation	0	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0	0	0
basal cell carcinoma	14	0	1	0	55	9	0	0	0	0	4	0	4	0	5	0	10	1
dermatofibroma	2	0	0	0	2	12	0	0	0	2	3	0	1	0	0	0	1	0
lentigo NOS	2	0	0	4	0	0	3	0	0	2	0	0	0	0	0	1	2	0
lentigo simplex	0	0	0	1	0	0	2	2	0	0	3	0	0	0	0	0	1	0
lichenoid keratosis	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
melanoma	26	0	0	17	27	21	3	6	0	1.4e+02	62	0	58	0	13	4	57	2
nevus	55	0	1	41	1.4e+02	2.2e+02	10	41	1	3e+02	2.5e+03	7	1.4e+02	0	85	6	68	91
other	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
pigmented benign keratosis	43	0	0	1	26	27	0	0	0	16	12	0	66	0	1	0	32	2
scar	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
seborrheic keratosis	1	0	0	0	5	3	1	0	0	5	25	1	3	0	32	1	2	0
solar lentigo	0	0	0	0	2	0	1	1	0	2	0	0	0	0	0	3	1	0
squamous cell carcinoma	11	0	0	0	8	2	0	0	0	2	2	0	3	0	1	0	19	0
vascular lesion	0	0	0	0	1	1	0	0	0	0	4	0	1	0	0	0	0	[illegible]

Figure 8. Confusion Matrix after training with no class weights and sequence generator enabled. Trail: Seq-1

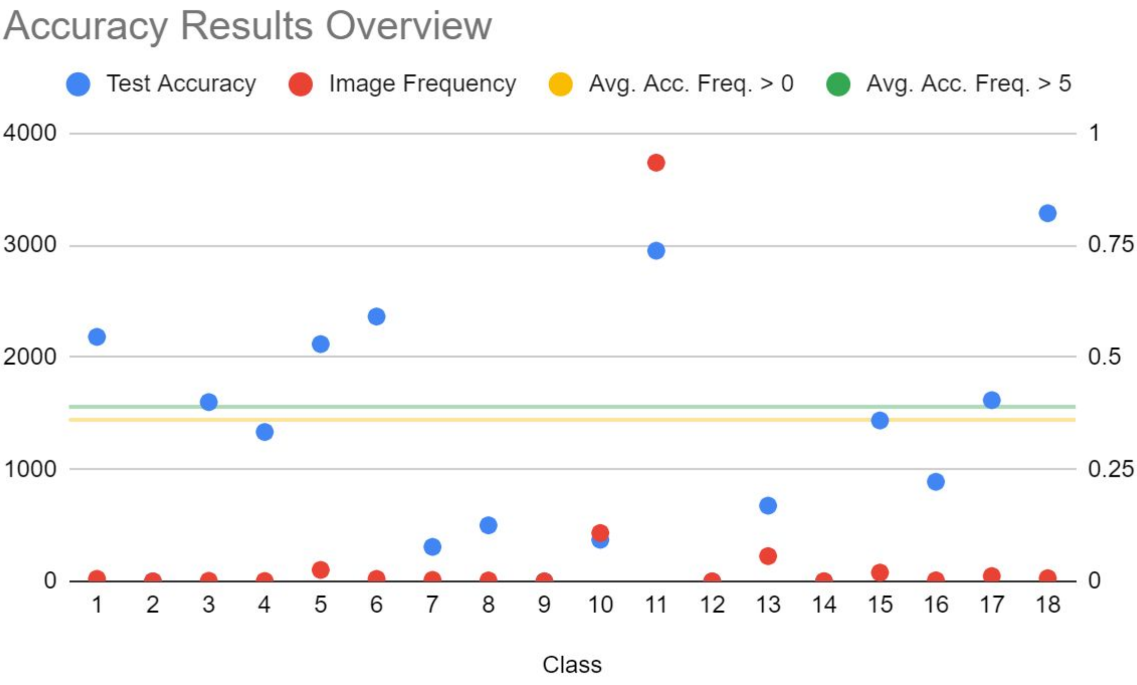


Figure 9. Model testing results for trail Seq-1 showing accuracy for different classes/diagnoses. The diagnoses are listed in the same order as the x axis for the confusion matrices. The yellow line shows the average accuracy of classes containing at least one test image. Green line shows average accuracy for the classes that contain over 5 test images. This is true for all “Accuracy Results Overview”-plots.

The confusion matrix above (Figure 8) shows some of the first results when using the sequence generator and not weighting the classes. The model was trained for approximately 20 epochs of the training data.

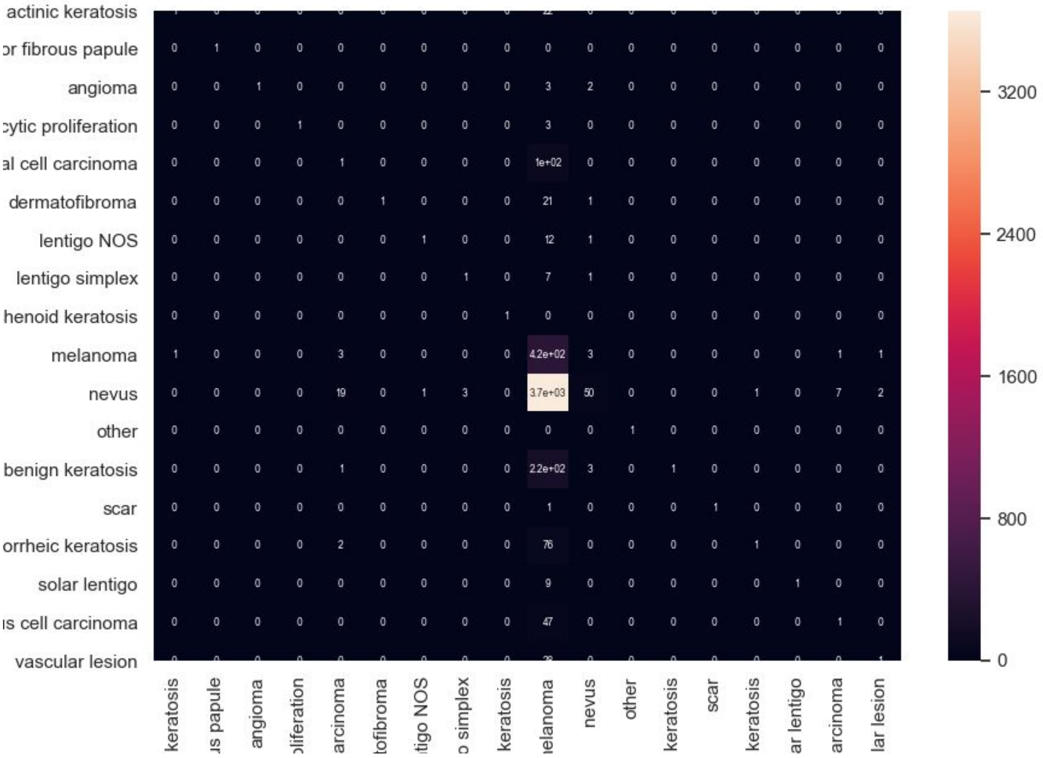


Figure 10. Confusion matrix showing testing results after 12 epochs of training with adjusted class weights and no sequence generator. Trail: Wei-1

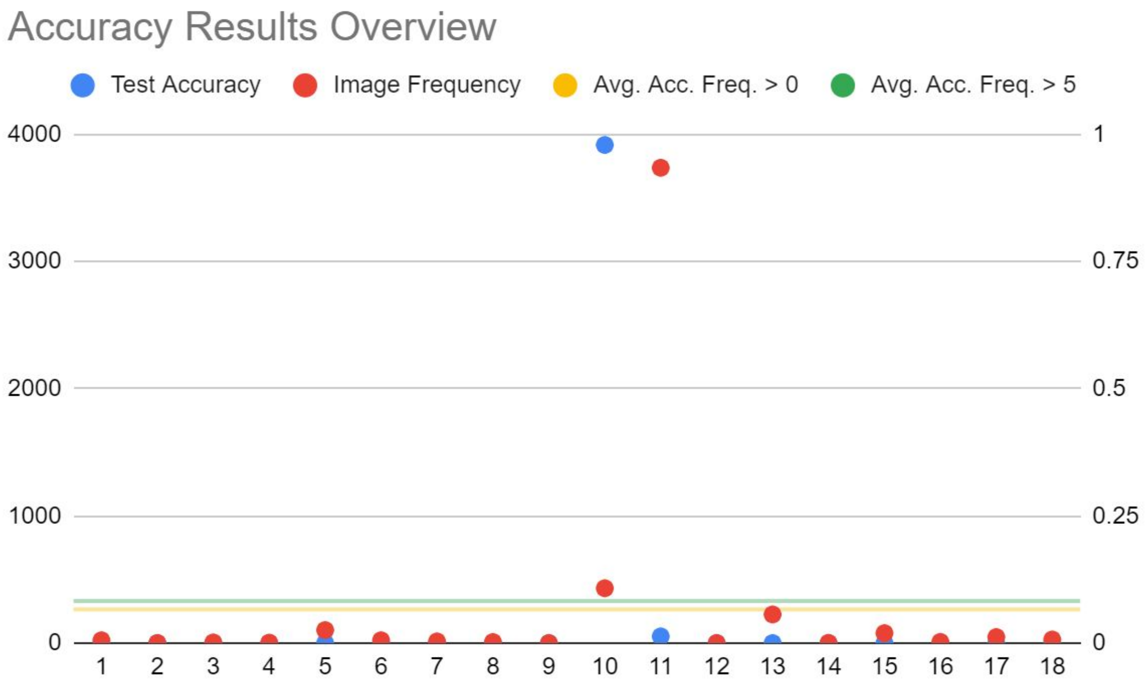
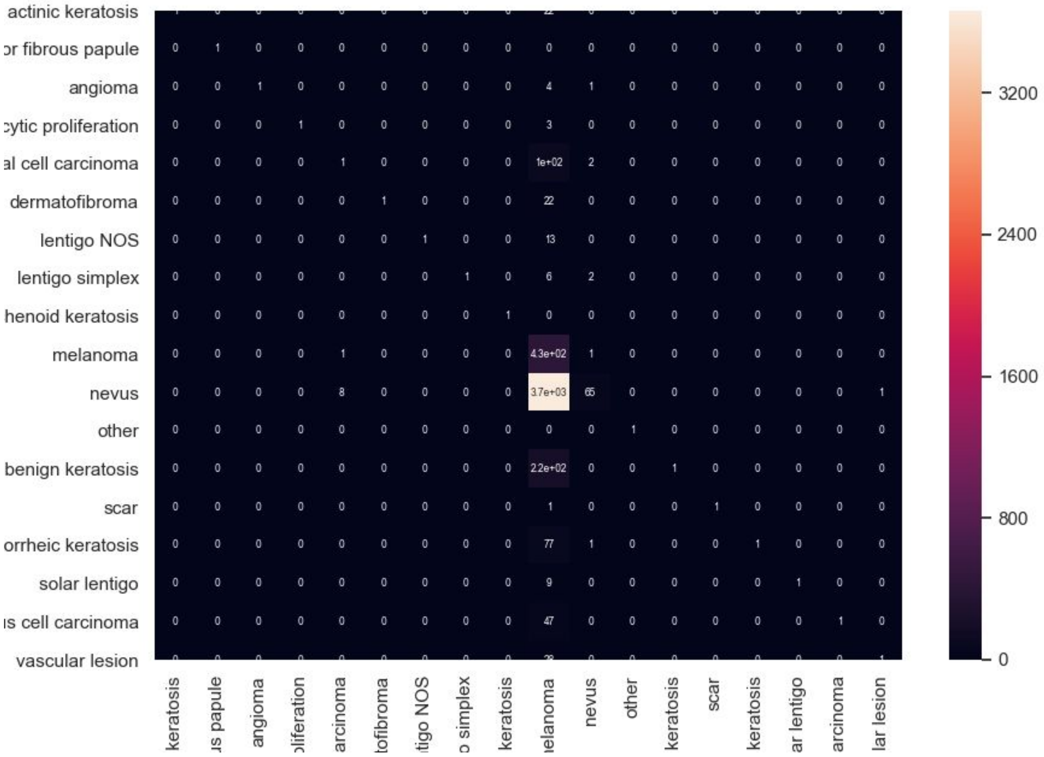


Figure 11. Model testing results for trail Wei-1 showing accuracy for different classes/diagnoses. Refer to figure 9 for an explanation of the metrics used in this figure.



	keratosis	ıs papule	angioma	ɔliferation	arcinoma	tofibroma	ıtigo NOS	ɔ simplex	keratosis	ıelanoma	nevus	other	keratosis	scar	keratosis	ar lentigo	arcinoma	lar lesion
actinic keratosis	1	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0
ɔr fibrous papule	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
angioma	0	0	1	0	0	0	0	0	0	4	1	0	0	0	0	0	0	0
ɔytic proliferation	0	0	0	1	0	0	0	0	0	3	0	0	0	0	0	0	0	0
al cell carcinoma	0	0	0	0	1	0	0	0	0	1e+02	2	0	0	0	0	0	0	0
dermatofibroma	0	0	0	0	0	1	0	0	0	22	0	0	0	0	0	0	0	0
lentigo NOS	0	0	0	0	0	0	1	0	0	13	0	0	0	0	0	0	0	0
lentigo simplex	0	0	0	0	0	0	0	1	0	6	2	0	0	0	0	0	0	0
henoid keratosis	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
melanoma	0	0	0	0	1	0	0	0	0	4.3e+02	1	0	0	0	0	0	0	0
nevus	0	0	0	0	8	0	0	0	0	3.7e+03	65	0	0	0	0	0	0	1
other	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
benign keratosis	0	0	0	0	0	0	0	0	0	2.2e+02	0	0	1	0	0	0	0	0
scar	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
orrheic keratosis	0	0	0	0	0	0	0	0	0	77	1	0	0	0	1	0	0	0
solar lentigo	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	1	0	0
ıs cell carcinoma	0	0	0	0	0	0	0	0	0	47	0	0	0	0	0	0	1	0
vascular lesion	0	0	0	0	0	0	0	0	0	[illegible]	0	0	0	0	0	0	0	1

Figure 12. Confusion matrix showing results after 70 additional epochs with adjusted class weights and no sequence generator. Trail: Wei-2

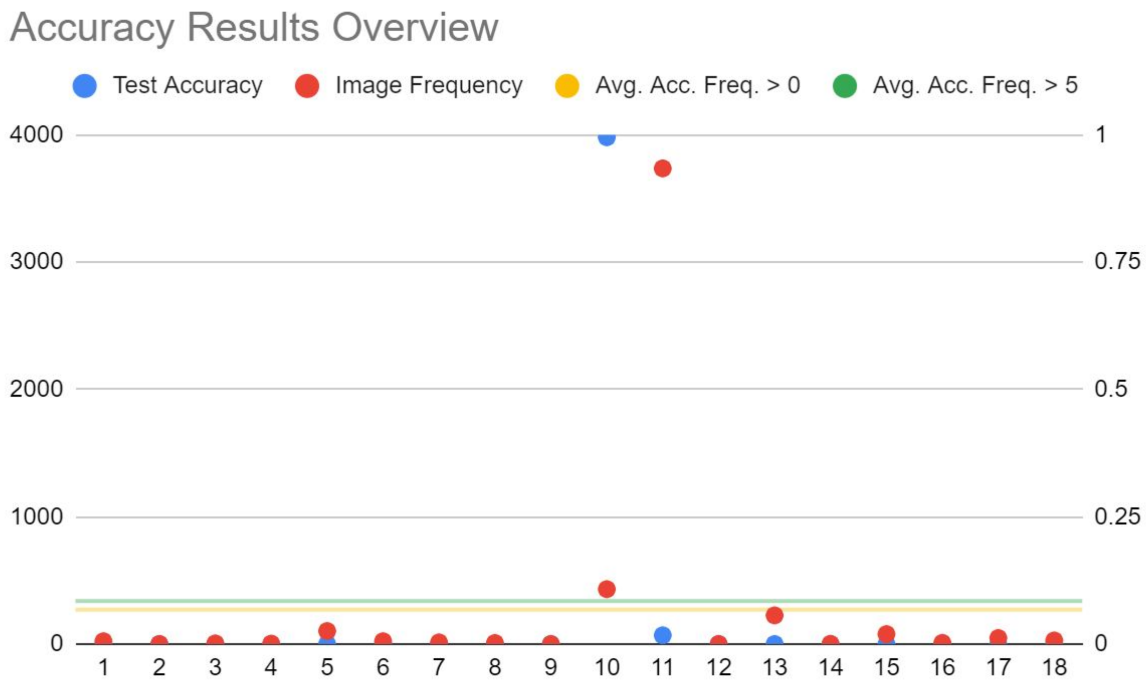


Figure 13. Model testing results for trail Wei-2 showing accuracy for different

classes/diagnoses. Refer to figure 9 for an explanation of the metrics used in this figure.

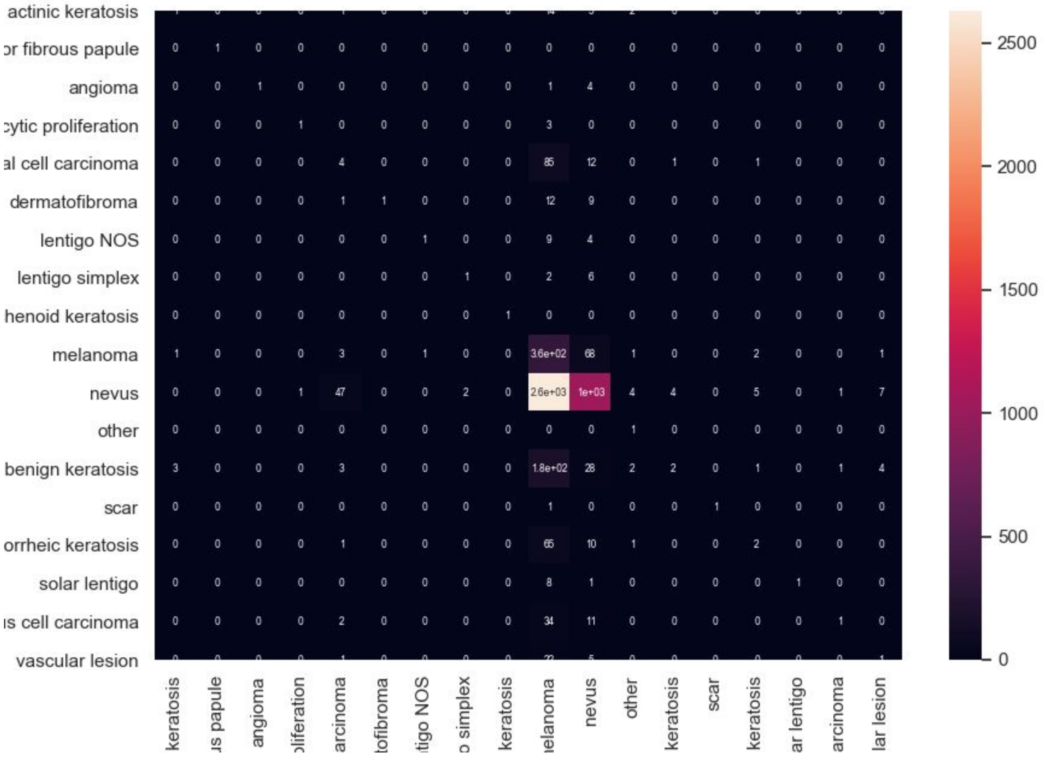


Figure 14. Confusion matrix showing epoch with best validation score during training with adjusted weights. Trail: Wei-3

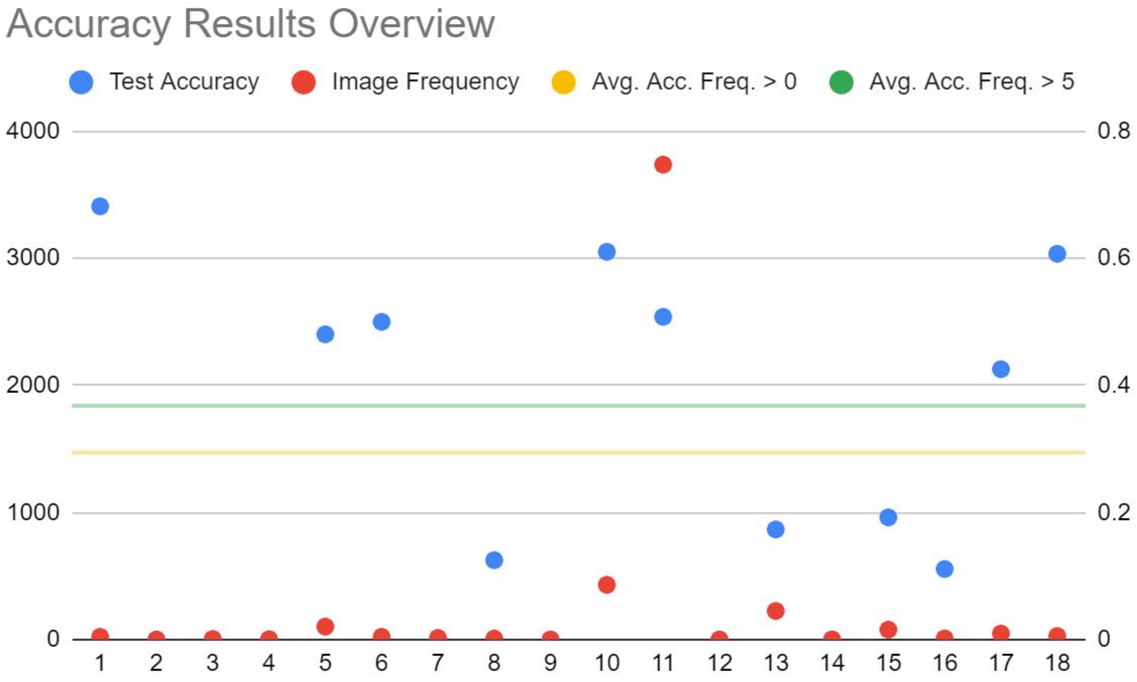


Figure 15. Model testing results for trail Wei-3 showing accuracy for different classes/diagnoses. Refer to figure 9 for an explanation of the metrics used in this figure.

As seen in the confusion matrix heatmap in figure 14, the model for trail Wei-3 is biased toward melanoma as a result of weights adjustments. The majority of test images where the true class is nevus were predicted as melanoma. Sensitivity for melanoma is approximately 84% and specificity is 28%.

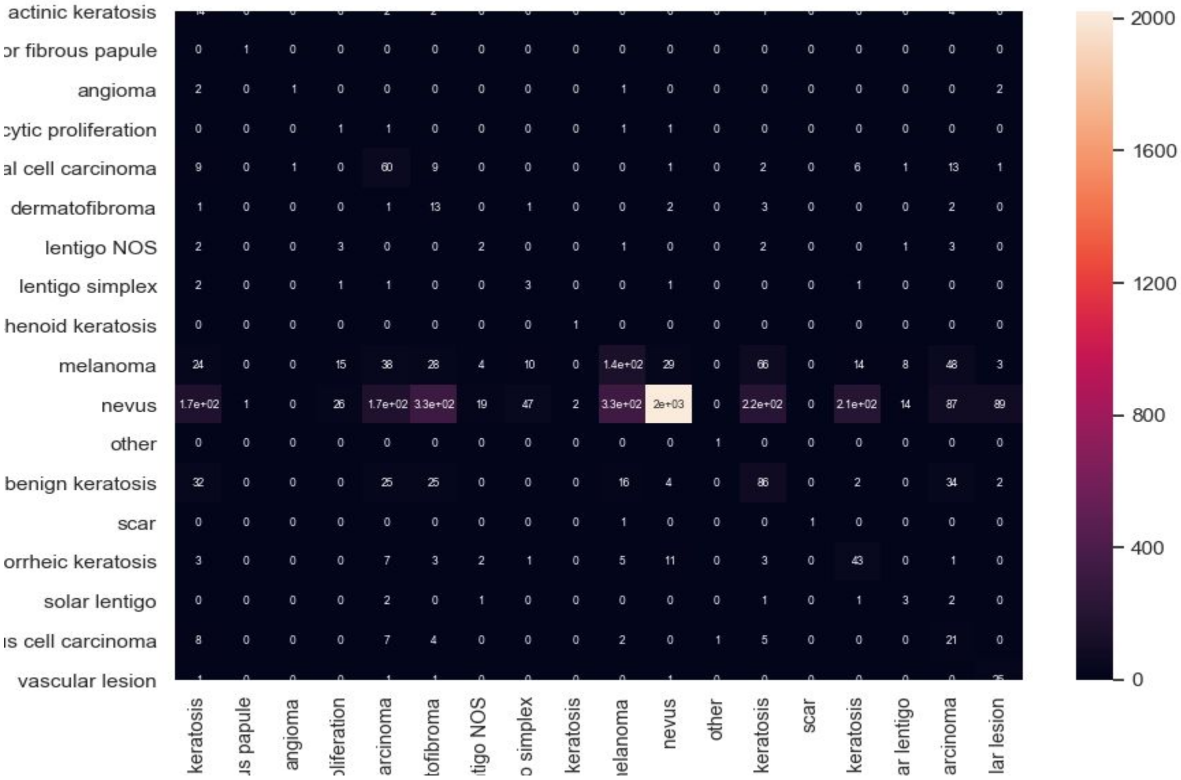


Figure 16. Confusion matrix showing test results after 70 epochs with sequence generator and no weights. Trail: Seq-2

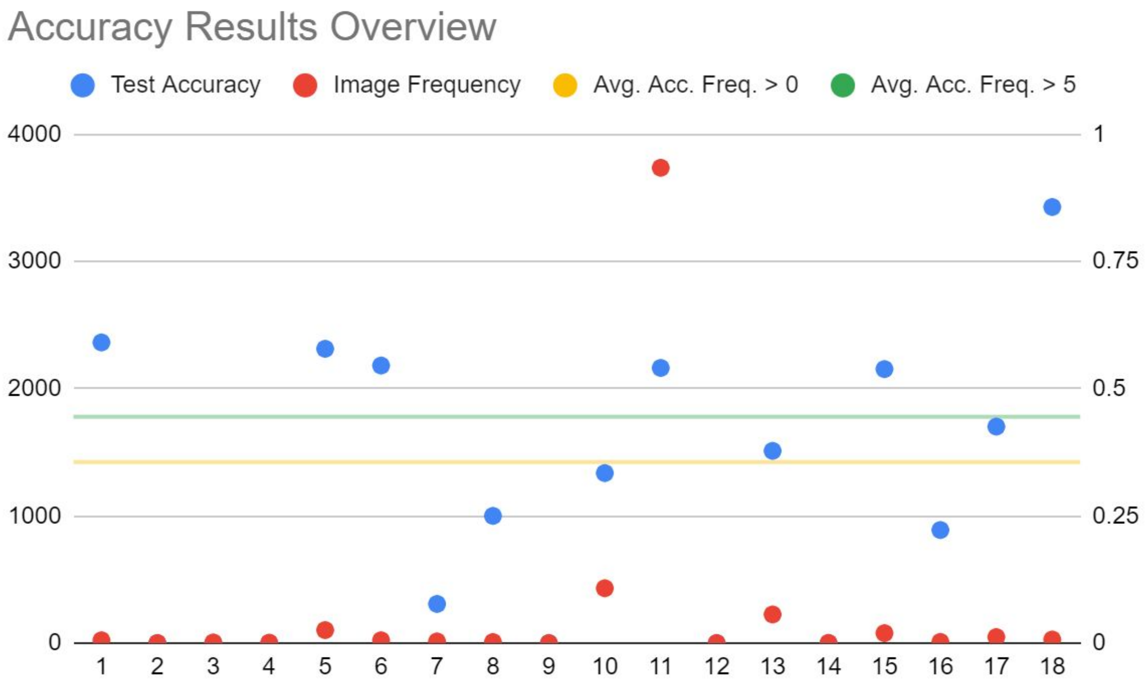


Figure 17. Model testing results for trail Seq-2 showing accuracy for different classes/diagnoses. Refer to figure 9 for an explanation of the metrics used in this figure.

This trail (Seq-2) has a weak diagonal line from top left to bottom right on the confusion matrix. Increased accuracy is seen in different types of keratosis and carcinoma. When the model is fed nevus many instances are predicted incorrectly. The sensitivity and specificity for melanoma is around 32% and 90% respectively.

Overall the top 3 trials in terms of average accuracy was Seq-2, Seq-1 and Wei-3. These achieved an average accuracy for categories with at least one testing image of 35.6%, 36.0% and 29.4% respectively. In terms of accuracy where more than 5 test images are available, Seq-2 achieved an accuracy of 44.5%, Seq-1 39.0% and Wei-3 36.8%. Wei-3 is more sensitive to melanoma with a sensitivity of 84%. The sensitivity of Seq-2 is 32%.

It is important to note that the *Avg. Acc. > 0* average accuracy was taken over 15 classes since 3 classes contain zero images. This means that if the model were to make random decisions, it would result in an average accuracy of around 7%. In this case the best model was able to outperform random decisions by a factor of 5.4 or 29 percentile units. For the case of classes containing more than five test images the random choice accuracy would be 8.3%. The trail Seq-2 was able to outperform 8.3% by a factor of 5.34 or 36 percentile units.

5. Conclusion

To answer the main question of this study, how accurately a machine learning model can diagnose skin lesions if developed using free and easily available information, the best model Seq-2 was able to achieve an average accuracy of 36% on the test data where empty test categories were removed. It achieved an average accuracy of 45% for all categories that contained over 5 images. Since it was able to significantly outperform random decision making, it could be considered a success in the way that it gives an indication of the right diagnosis.

6. Discussion

When it comes to the accuracy of the software, although it was better than random decision making, test results should be compared to real world statistics in order to evaluate the practical performance of the models. The performance of trail Seq-2 and Wei-3 will be compared to statistics for dermatologists abilities to identify melanoma. A Study review evaluated the accuracy of dermatologists for different scenarios. One of these scenarios was when dermatologists performed an in-person diagnosis for 5567 cases of suspected melanoma. When dermatologists were presented with skin lesions where melanoma was present (confirmed later by biopsy), they predicted positive for melanoma in approximately 92% of melanoma positive cases (sensitivity). On the other hand, when presented melanoma negative skin lesions, around 80% were correctly predicted as negative (specificity). This indicates a slight bias towards predicting true for melanoma amongst dermatologists in this scenario. A likely reason is that it is safer to predict true for melanoma where none is present, than it is to predict false for melanoma when it is present. The later could potentially result in death. Also, the dermatologists knew that melanoma was suspected in the skin lesion and a small factor for the bias could be a tendency for the dermatologists to agree with the proposed suspicion. When dermatologists from this review performed diagnostics based on images, rather than in-person, accuracy was stated to be much lower. However, numbers were not provided. [3]

Another comparison could be made to the performance of modern diagnostics software for skin lesions. A study by a group at Stanford University has achieved state-of-the-art performance, with greater sensitivity and specificity than board-certified dermatologists. The results from said study will be used for a comparison by selecting points from the presented graphs for sensitivity and specificity for melanoma detection. [4]

[3] Dinnes J, Deeks JJ, Grainge MJ, Chuchu N, Ferrante di Ruffano L, Matin RN, Thomson DR, Wong K, Aldridge R, Abbott R, Fawzy M, Bayliss SE, Takwoingi Y, Davenport C, Godfrey K, Walter FM, Williams HC, *How accurate is visual inspection of skin lesions with the naked eye for diagnosis of melanoma in adults?,* 2018

[4] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun, *Dermatologist-level Classification of Skin Cancer,* 2017, https://www.nature.com/articles/nature21056.epdf?author_access_token=8oxIcYWf5UNrNpHsUHd2S

	Sensitivity	Specificity
Average dermatologist, in-person diagnosis	92%	80%
State of the art AI at 92% sens.	92%	92%
State of the art AI at 80% spec.	96%	80%
Seq-2, image-based diagnosis	32%	90%
Wei-3, image-based diagnosis	84%	28%

Figure 19. Table showing the sensitivity and specificity of different diagnostic methods.

When inspecting the results, figure 19 shows that the machine learning model developed by the stanford team has the highest overall performance. This is followed by the performance of average dermatologists reported by the study review, and lastly the machine learning models developed for this study. The differences in performance depends on many factors such as the amount of categories used for classification, the processing power and data available for machine learning models. Also, the amount of time spent on developing the software. The dermatologists performs binary classification thereby having a slight advantage over the machine learning models. Furthermore, the stanford models use a larger dataset not easily available to the public and a larger development team. Also, the team has access to more computing power which results in faster training of the models.

In order to achieve better results for the models in this study, some modifications could be made to the method. For instance, types of skin lesions with similar characteristics and risk levels could be grouped together. By having a larger amount of data in each category, the performance could be evaluated more easily and it might result in enhanced ability for the neural network to generalize a pattern between appearance and risk level. A larger dataset, more varied data-augmentation, and longer training time could help performance.

When it comes to the CNN (convolutional neural network) itself, a higher resolution of input images could allow for more detail to be obtained. Also the architecture could be changed from VGG16 to a modern architecture like *EfficientNet* or *ResNet.*[5,6] By implementing all of the previously mentioned modifications, performance and accuracy will likely be significantly improved.

tRgN0jAjWel9jnR3ZoTv0NXpMHRAJy8Qn10ys2O4tuPakXos4UhQAFZ750CsBNMMsISFHIKinKDMKjShCpHIlYPYUHhNzkn6pSnOCt0Ftf6

[5] Alexander Kolesnikov • Lucas Beyer • Xiaohua Zhai • Joan Puigcerver • Jessica Yung • Sylvain Gelly • Neil Houlsby, *Large Scale Learning of General Visual Representations for Transfer*, 2019, https://paperswithcode.com/paper/large-scale-learning-of-general-visual

[6] Qizhe Xie • Minh-Thang Luong • Eduard Hovy • Quoc V. Le, *Self-training with Noisy Student improves ImageNet classification*, 2019, https://paperswithcode.com/paper/self-training-with-noisy-student-improves

The validity of the data used in this project is solid. Dermoscopic images from the ISIC Archive have been collected through several databases, each database containing images provided by different dermatologists. The result is varied data that has been shuffled in this study before separating into training, validation and test datasets. This ensures the models are not optimized for a specific source of data, but are rather more general and ready for real world applications. In terms of reliability however, some categories of the test data contain few images, resulting in high inaccuracy during evaluation. Even though this has been accounted for by not including categories with fewer than 6 images in some average calculations, it would be better to combine similar categories as mentioned previously in order to increase the number of test images in small categories. This modification would significantly increase reliability of test data evaluations.

To answer the question: *How accurately can a machine learning algorithm, developed using free, easily available information, diagnose skin lesions?*, it is relatively simple to achieve 32% sensitivity and 90% specificity using the method from this study. The application of this software could be for example an app that the average person can use when investigating a suspicious skin lesion on their body. Overall, the results serve the purpose of this paper by indicating that machine learning can in fact be used to diagnose skin lesions.

Reference List

Alexander Kolesnikov • Lucas Beyer • Xiaohua Zhai • Joan Puigcerver • Jessica Yung • Sylvain Gelly • Neil Houlsby, *Large Scale Learning of General Visual Representations for Transfer*, 2019, https://paperswithcode.com/paper/large-scale-learning-of-general-visual

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau & Sebastian Thrun, *Dermatologist-level Classification of Skin Cancer,* 2017, *https://www.nature.com/articles/nature21056.epdf?author_access_token=8oxIcYWf5UNrNpHsUHd2StRgN0jAjWel9jnR3ZoTv0NXpMHRAJy8Qn10ys2O4tuPakXos4UhQAFZ750CsBNMMsISFHIKinKDMKjShCpHIlYPYUHhNzkn6pSnOCt0Ftf6*

Dinnes J, Deeks JJ, Grainge MJ, Chuchu N, Ferrante di Ruffano L, Matin RN, Thomson DR, Wong K, Aldridge R, Abbott R, Fawzy M, Bayliss SE, Takwoingi Y, Davenport C, Godfrey K, Walter FM, Williams HC, *How accurate is visual inspection of skin lesions with the naked eye for diagnosis of melanoma in adults?,* 2018

International Skin Imaging Collaboration, *ISIC Archive,* 2019, https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main

Karen Simonyan and Andrew Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2015, https://arxiv.org/pdf/1409.1556.pdf

Qizhe Xie • Minh-Thang Luong • Eduard Hovy • Quoc V. Le, *Self-training with Noisy Student improves ImageNet classification*, 2019, https://paperswithcode.com/paper/self-training-with-noisy-student-improves