Dermatological Image Classification In Detecting Melanoma

Axel Blom
M.Sc. Data Science & AI
DAT341, Group 36
axelblo@chalmers.se

Andreas Helgesson
M.Sc. Data Science & AI
DAT341, Group 36
andhelge@chalmers.se

Johanna Norell
M.Sc. Data Science & AI
DAT341, Group 36
jnorell@chalmers.se

Abstract

By applying transfer learning using a pre-trained image classifier model, skin marks can be classified as being cancerous melanoma, or benign nevus, with an accuracy of 83.3% on validation data. When applied to a blind test set, an accuracy of 84.1% was achieved.

1 Introduction

In the analysis of skin marks to detect the presence of cancerous melanoma or benign nevus, artificial intelligence and particularly Convolutional Neural Networks (CNNs), which categorizes images by pattern recognition, has demonstrated significant utility. As skin cancer has arisen as one of the most prevalent cancers globally, employing CNNs could possibly automate a labor-intensive task for medical professionals. Furthermore, it has the potential to surpass human diagnostic accuracy with adequate training. This report investigates various machine learning architectures and tools, assessing their efficacy in melanoma detection.

2 Preparing the Data

It was quickly found that training a large CNN model on the entire dataset would take too much time to efficiently try multiple configurations in a reasonable time frame for this project. Therefore, to complete the initial search for the best model configuration withing a reasonable time, a smaller subset was created from the original training dataset. This subset contained 100 images for each class, compared to over 3000 in the complete dataset.

After completing the search for the best model on this subset, the full data set was used to final evaluation.

3 CNNs for Image Classification

In this section the different model and their related configuration will be shown. A motivation to the choice of model parameters will also be presented.

3.1 Model parameters

For this task, we were given a balanced dataset to train our model, for which the accuracy score was used to evaluate its performance. More commonly in medical image classification, the F1-score is preferred because datasets tend to be imbalanced, making accuracy a less reliable measure. Given the balanced nature of our dataset, however, the accuracy score was deemed an appropriate metric for evaluation.

The layer structure of all models was kept consistent and consists of two convolutional layers followed by max pooling and ReLU activation. Further transiting into two fully connected layers with a dropout layer introduced after the first fully connected layer to reduce overfitting. This design aims to balances complexity and efficiency.

Moving on, cross entropy loss was selected as the loss function as its ability to penalize incorrect classifications more heavily makes it useful for enhancing the model's accuracy in medical image classification.

The epoch count for each model was carefully calibrated in each separate case. Initially set at 10 epochs, adjustments were made based on an analysis of the performance graphs for both the training and validation data, allowing us to identify the onset of overfitting.

3.2 Establishing a Baseline

When establishing the baseline model, it needed careful tuning to prevent both underfitting and overfitting. Initially, underfitting was prominent; however, increasing the number of layers and epochs subsequently led to overfitting. Intially a model with four layers was established, but lead to immense overfitting after only one epoch. Downgrading from four to two layers and significantly reducing the number of nodes, from 512 in the last hidden layer to 16, the model started to display reasonable accuracy.

Through iterative refinement, with the architecture described in section 3.1, balanced model was achieved. For this baseline, the number of epochs was set to four to prevent overfitting, resulting in an accuracy score of 81.2%.

3.3 Batch Normalization

With the help of batch normalization, one can achieve a high accuracy without necessarily needing more training steps. Batch normalization improves the mathematical stability and allows us to use higher learning rates and reduces the models sensitivity to initialization (Ioffe and Szegedy, 2015).

By introducing two layers of batch normalization following the hidden layers in the baseline model, and limiting the model to a single epoch due to rapid onset of overfitting, an accuracy of 78.2% was achieved on the validation data.

3.4 Residual Connections

As CNN's grow deeper the problem of vanishing and exploding gradients occur which makes the computations unstable. To reduce this issue residual connection which bypasses some of the layers in the neural network and creates a direct connection between the input and the output of the network. By doing so, the gradients become more stable which enables deeper models to be built. (He et al., 2016).

Subsequently, a new model incorporating a residual connection, but no batch normalization, was developed to assess improvements over the baseline classifier. This model exhibited consistent performance improvement across all 10 epochs, showing no indications of overfitting. Therefore the number of epochs was increased to 20 epochs. Ultimately, the epoch count was optimized to 16, achieving an accuracy of 80.4% on the validation data.

3.5 Data Augmentation

Data Augmentation can be used to improve robustness in several different machine learning algorithms, and is significantly useful in image classification because of the simple techniques to achieve augmented data close to realistic variations in the data.

The purpose of data augmentation is to introduce variance through realistic images. It is crucial to be selective about the features allowed to change during the augmentation process. In this task, variations in rotation, brightness, and contrast were introduced. However, variations in saturation and cropping were deemed to potentially alter the images' characteristics significantly, so no variance in these features was permitted.

When data augmentation was introduced to the baseline classifier the best performance without overfitting was received after two epochs, with an accuracy score of 68.5% on the validation data. Note that this is a drastically lower than previously tested configurations.

3.6 Transfer learning

To evaluate the effect of transfer learning, AlexNet was used. The choice of AlexNet was natural in this assignment due to the lack of time; AlexNet is reportedly much faster to train compared to other established networks (Anwar, 2019).

The AlexNet model used was pre-trained on the ImageNet database, which consists of more than 14 million images in a wide range of categories such as objects, animals and food [1]. This model was imported and trained for additional epochs on the entire training dataset, transferring its knowledge of image classification and feature extraction to this new data, after which it was found that 7 epoch gave the highest accuracy without resulting in much overfitting. This trained model reached an accuracy score of 83.3% on the validation data.

4 Evaluation on Validation Data

The performance of the difference models is visualized in figure 1.

Note that the model with residual connections was trained for 20 epochs as it did not show a clear plateau in accuracy in just 10 epochs, while all other models were trained for 10 epochs.

The model using transfer learning from AlexNet, which achieved the best performance, was selected for evaluating the blind test dataset. Further iterations were conducted to determine the optimal number of epochs more carefully, as the trend over 10 epochs was continuously positive. Ultimately, seven epochs were identified as the

[1] https://www.image-net.org/about.php

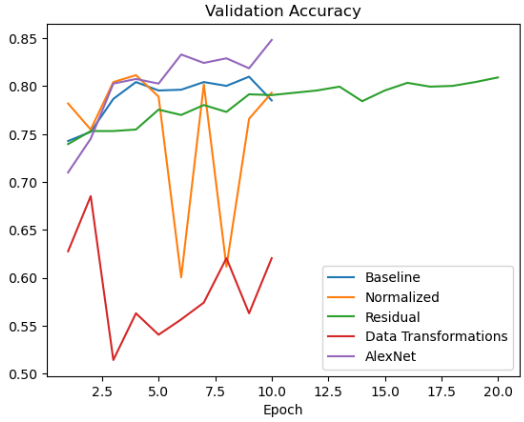

Figure 1: Validation accuracy over epochs for all models evaluated.

ideal duration to attain peak accuracy without signs of overfitting. See figure 2.

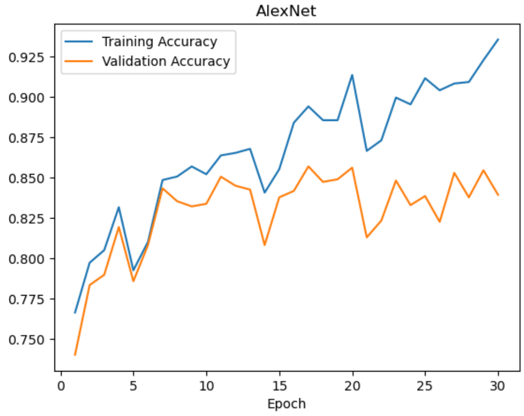

Figure 2: Accuracy score per epoch on training and validation data for the AlexNet model.

Evaluating the model on the validation data achieved an accuracy score of 84.3%, and these results are visalized in a confusion matrix in figure 3.

As per the confusion matrix, the precision derived is 83.2%, and the recall is 86.1%.

Lastly, evaluating this model on the blind test set gave an accuracy score 84.1%.

5 Discussion

The task to establish a CNN model for image classification of skin marks yielded several interesting obesrvation. Notably, one of the most important aspects is that most adaptions of the model, such as normalization, residual connections and data augmentation reduced the accuracy compared to the baseline case. The only exception was using transfer learning with AlexNet, which achieved 84.3% accuracy in comparison to the baseline accuracy score of 81.2%.

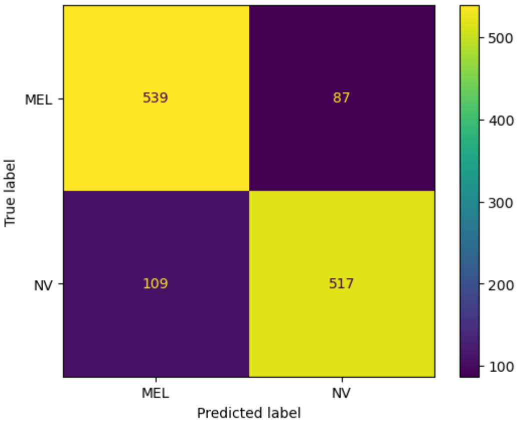

Figure 3: Confusion matrix for best model on validation data.

Remarkably, the application of transfer learning with AlexNet yielded a higher score than expected. Given that this model was initially trained on a broad range of general images, one might question its suitability for highly specialized tasks like medical image classification. The presumption could be that features derived from such diverse images would translate poorly to the specifics of medical diagnostics. However, contrary to these expectations, the application of AlexNet has led to the best performance in this task. This outcome challenged our assumptions about the transferability of features from general to specialized domains, indicating a potentially greater utility of generalized models in specific contexts than previously believed.

If analysing the best achieved accuracy itself, it underscores an important question of whether it is good enough to be used in a real world application. Using this model to classify skin marks, more than 15% of the patients would get an incorrect diagnosis. To further put the accuracy score into the context of the task, it is also interesting to derive more evaluation metrics from the confusion matrix. This issue will be discussed further in section 8.

5.1 Evaluation measurement

Selecting accuracy as the sole evaluation metric for our model's performance may seem simplistic for such a nuanced task. The choice of metric should align with the model's primary objective. For instance, in early cancer screening, minimizing false negatives is crucial to avoid overlooking treatable cases. Conversely, in later diagnostic stages, reducing false positives becomes important to prevent unnecessary treatment of healthy individuals. We found early detection the most probable objective of this model, but despite this we opted for accuracy due to its simplicity. A deeper understanding of the model's application could guide a more nuanced choice of evaluation metric.

In medical image classification, two crucial evaluation metrics are sensitivity and specificity. Sensitivity measures the proportion of actual positives correctly identified, in this context, cancerous skin marks. Specificity, on the other hand, assesses the accuracy in identifying healthy skin marks. For this application, the sensitivity and specificity scores are 86.1% and 82.6%, respectively. While high values for both metrics are desirable, the primary aim of deploying this model is likely early cancer detection. From this perspective, the false negative rate, calculated as $1 - sensitivity$, is 13.9%. This indicates that more than one in ten cancerous cases could be missed by the model, a concern that suggests the model's performance may not be sufficiently reliable for its intended use.

6 Conclusion

Of the model CNN configurations tested, only transfer learning provided an increase in accuracy compared to the baseline case.

However, the accuracy score of 84.3% on the validation data means that more than one in ten of all cancerous cases would be missed by this optimal model, creating doubts as to whether it is sufficiently reliable for use in an actual medical environment.

7 Limitations

One limiting factor is the data that was collected. Firstly, the quality of a picture is highly dependent on the equipment used and the lighting when the picture is taken. The training and test data were collected during similar circumstances and hence are of very similar quality. A real-world application would have pictures that are taken during different circumstances and the model would likely perform worse. To combat this drawback, data augmentation was used to try to get more variation in the data. The data augmentation resulted in worse performance, which is to be expected. However, data augmentation will probably have a positive impact on real-world applications. Our ”best model” on the test set did not use data augmentation, which limits its ability to generalize. A professional dermatologist's knowledge would also be useful to provide insights about what features to allow changes on, when performing the data augmentation. The data set also caries limitations in the representation of different skin tones, which is discussed further in section 8.

Another limitation of the data set is that it only has image data, our model does not take size into account for instance. Furthermore, it does not take into account any symptoms such as itchiness or bleeding, and no blood values or other similar data points. It is very possible that the use of the mentioned more traditional data points, in conjunction with the image analysis, would lead to a higher accuracy in the predictions.

In the construction of the neural network a lot of trial and error was used, this required us to use a smaller training data set to be able to test different possible configurations. which could lead to randomness in the smaller training set impacting the evaluated accuracy, and or that the model needed more data to be able to train to higher accuracy levels. This was done due to time limitations. The final number of epochs was tuned on the full training set using the chosen neural network construction. Fine-tuning the epochs for different configurations with the full data set could have led to some insights. For instance: the residual connections model had lower accuracy but was much more stable with increasing epochs. It would also be interesting to look into different combinations of the tested method, or some type of enable method. An ensemble of multiple of these solutions could be a good idea, as we could get the accuracy of AlexNet and the stability of residual connection models. Finally, some more time could have been spent trying to find a more task-specific pre-trained model, as AlexNet is a very large and general model. If found, a pre-trained model made for categorizing skin type would surely show promise if tested.

8 Ethical considerations

As all design choices in machine learning will affect the perceived result or success of the reader of this report, a discussion about the ethical considerations related to this assignment will follow. The section will discuss fairness, error functions, misleading claims about performance and limitations of the training data.

8.1 Fairness and error functions

Considering fairness in model design involves examining how our methodologies and reporting impact performance measurements (Lindholm et al., 2022). A critical aspect of assessing model fairness is analyzing its performance across different subcategories within the data. Although the data in this study were not initially categorized, a logical division might be based on skin tone. Such a categorization warrants a detailed analysis to evaluate the model's performance across these groups, which is crucial for assessing fairness.

This approach shifts the fairness discussion towards evaluation metrics, particularly the selection of an evaluation metrics. As previously mentioned, the accuracy score alone may not adequately capture the model's performance nuances, especially concerning fairness. Instead, metrics like the false negative rate or the false positive rate should be considered, depending on how the model is deployed.

Crucially, examining the false negative rate across subcategories, such as between light- and dark-skinned individuals can reveal potential biases. Any significant discrepancy, where one subcategory experiences a higher false negative rate, indicates discrimination. In contexts like Sweden, where equality in healthcare access is legally mandated, such differences could jeopardize laws ensuring equal healthcare rights. Therefore, to address these aspects to prevent any form of unequal treatment or healthcare delivery is very important.

8.2 Misleading Claims about Performance

The discussion around evaluation metrics, particularly accuracy versus sensitivity and precision, underscores the potential for misleading representations when the primary aim is to highlight favorable outcomes. In tasks related to real-world problems, there might be secondary objectives, such as promoting a product or advancing a research career, rather than contributing meaningfully to the field.

Ideally, this task should contextualize the model's results by comparing them with alternative approaches to addressing the same problem. Evaluating the performance of our most effective CNN not only against simpler machine learning algorithms, like linear regression, but also against traditional methods, such as manual examinations by medical professionals, would provide a more grounded assessment of the model's efficiency.

Moreover, it's crucial to ensure that claims about the model's performance are not only accurate but also interpretive. Shifting from solely presenting accuracy scores to a more nuanced discussion that includes sensitivity and specificity, explicitly defined within our task's context, greatly reduces the risk of misinterpretation. By doing so, we advocate for a more transparent and honest evaluation of our model, ensuring that its purported benefits are both real and relevant to the intended application.

8.3 Limitations of the data

Data quality and diversity are foundational to the performance of machine learning algorithms. In the context of this task, the model's performance is primarily limited by the representativeness of the data set. The majority of images used are of light-skinned individuals, reflecting the demographics of the countries where the data originated. This can lead to a biased model, as it may not perform equally well for other skin tones. It is crucial to audit the data carefully, as the model can only replicate learned relationships without understanding their real-world implications (Lindholm et al., 2022).

References

A. Anwar. 2019. Difference between alexnet, vggnet, resnet, and inception. https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96. Accessed: 2024-02-21.

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. https://arxiv.org/pdf/1512.03385.pdf. Accessed: 2024-02-28.

S. Ioffe and C. Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. https://arxiv.org/abs/1502.03167. Accessed: 2024-02-28.

A. Lindholm, N. Wahlström, F. Lindsten, and T. Schön. 2022. Machine learning - a first course for engineers and scientists. http://smlbook.org/book/sml-book-draft-latest.pdf. Accessed: 2024-03-06.